©MARLIN E. RICE

Next Generation Sequencing for Invertebrate Virus Discovery

-a practical approach

Sijun Liu & Bryony C. Bonning

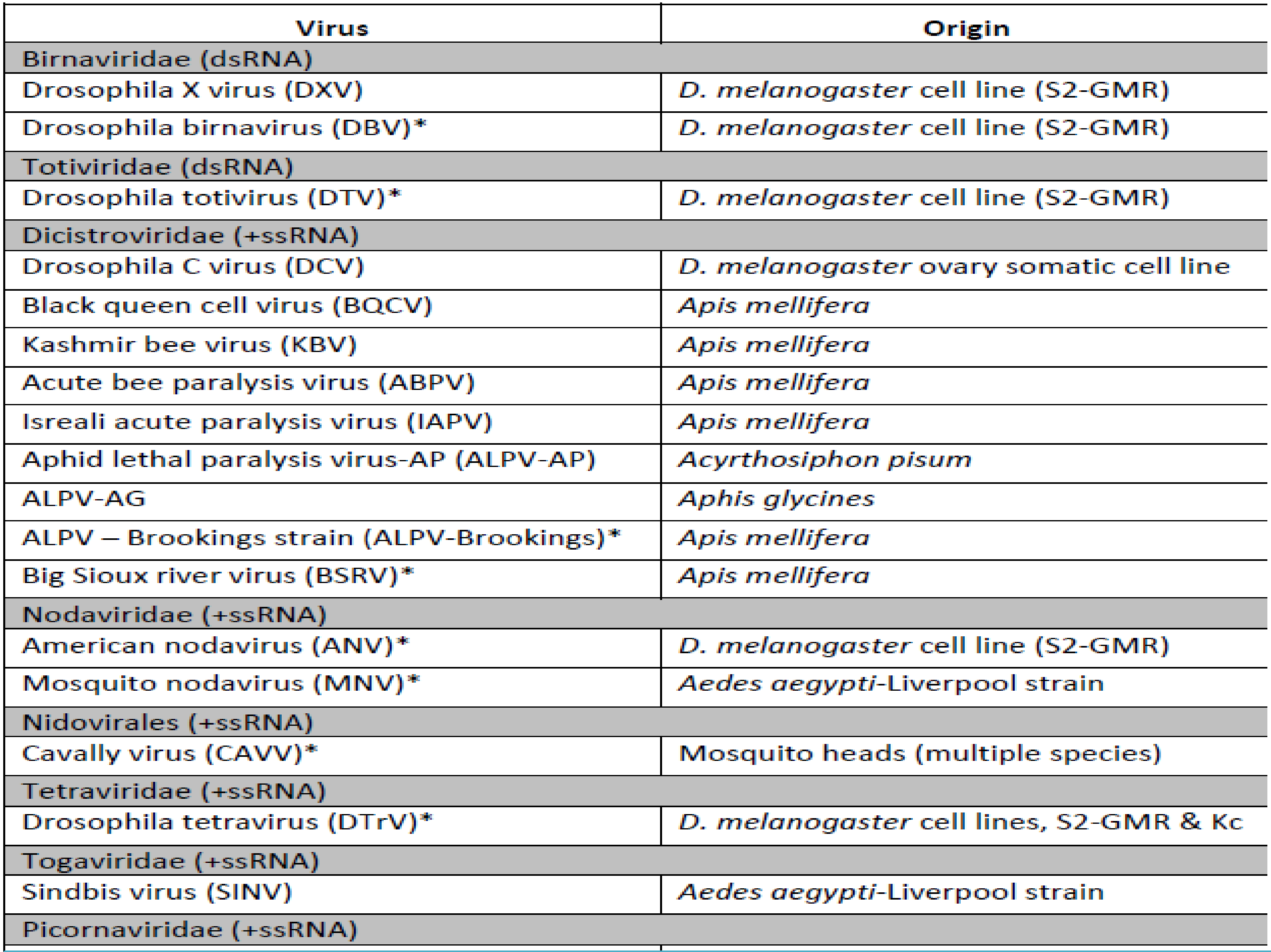
Iowa State University, USA

8-14-2013 SIP Pittsburgh

Outline

- Introduction: Why use NGS?
  - Traditional approach for virus discovery
  - Next Generation Sequencing (NGS)
  - Advantages of NGS for virus discovery
- How it's done
  - Sample selection
  - Sequencing library preparation
  - Sequencing method
  - Assembly of sequencing reads
  - Identification of viral sequence
  - Assembly of viral genome

Insect viruses detected/discovered by use of NGS

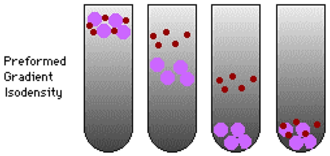


Virus	Origin
Birnaviridae (dsRNA)	
Drosophila X virus (DXV)	*D. melanogaster* cell line (S2-GMR)
Drosophila birnavirus (DBV)*	*D. melanogaster* cell line (S2-GMR)
Totiviridae (dsRNA)	
Drosophila totivirus (DTV)*	*D. melanogaster* cell line (S2-GMR)
Dicistroviridae (+ssRNA)	
Drosophila C virus (DCV)	*D. melanogaster* ovary somatic cell line
Black queen cell virus (BQCV)	*Apis mellifera*
Kashmir bee virus (KBV)	*Apis mellifera*
Acute bee paralysis virus (ABPV)	*Apis mellifera*
Isreali acute paralysis virus (IAPV)	*Apis mellifera*
Aphid lethal paralysis virus-AP (ALPV-AP)	*Acyrthosiphon pisum*
ALPV-AG	*Aphis glycines*
ALPV – Brookings strain (ALPV-Brookings)*	*Apis mellifera*
Big Sioux river virus (BSRV)*	*Apis mellifera*
Nodaviridae (+ssRNA)	
American nodavirus (ANV)*	*D. melanogaster* cell line (S2-GMR)
Mosquito nodavirus (MNV)*	*Aedes aegypti*-Liverpool strain
Nidovirales (+ssRNA)	
Cavally virus (CAVV)*	Mosquito heads (multiple species)
Tetraviridae (+ssRNA)	
Drosophila tetravirus (DTrV)*	*D. melanogaster* cell lines, S2-GMR & Kc
Togaviridae (+ssRNA)	
Sindbis virus (SINV)	*Aedes aegypti*-Liverpool strain
Picornaviridae (+ssRNA)	

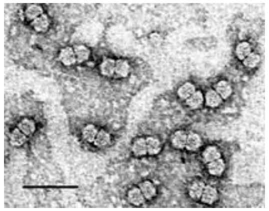
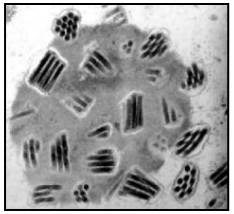
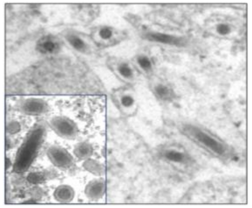
Liu S, Vijayendran D, Bonning BC. 2011. 3(10):1849-69.

Traditional Approach for Virus Discovery

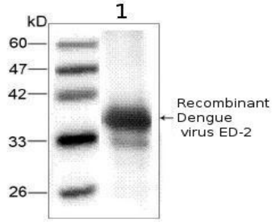
Collect samples that show disease symptoms

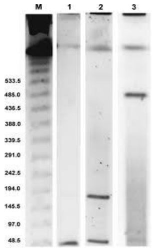


Isolate viruses

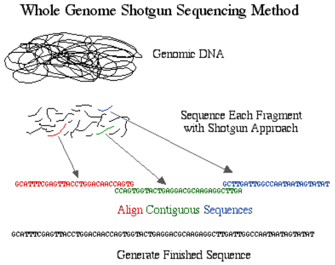
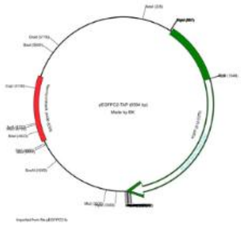
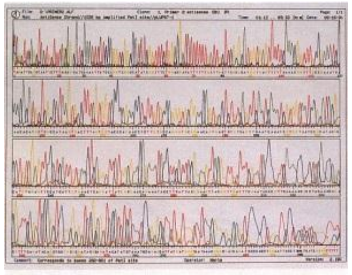




Observe virus particles





Identify viral genomes


Whole Genome Shotgun Sequencing Method
Genomic DNA
Sequence Each Fragment
with Shotgun Approach
GCATTTCGAGTTACCTGGACAACCAGTG
CCAGTGGTACTGAGGACGCAAGAGGCTTGA
GCTTGATTGGCCAATAATAGTATAT
Align Contiguous Sequences
GCATTTCGAGTTACCTGGACAACCAGTGGTACTGAGGACGCAAGAGGCTTGATTGGCCAATAATAGTATAT
Generate Finished Sequence



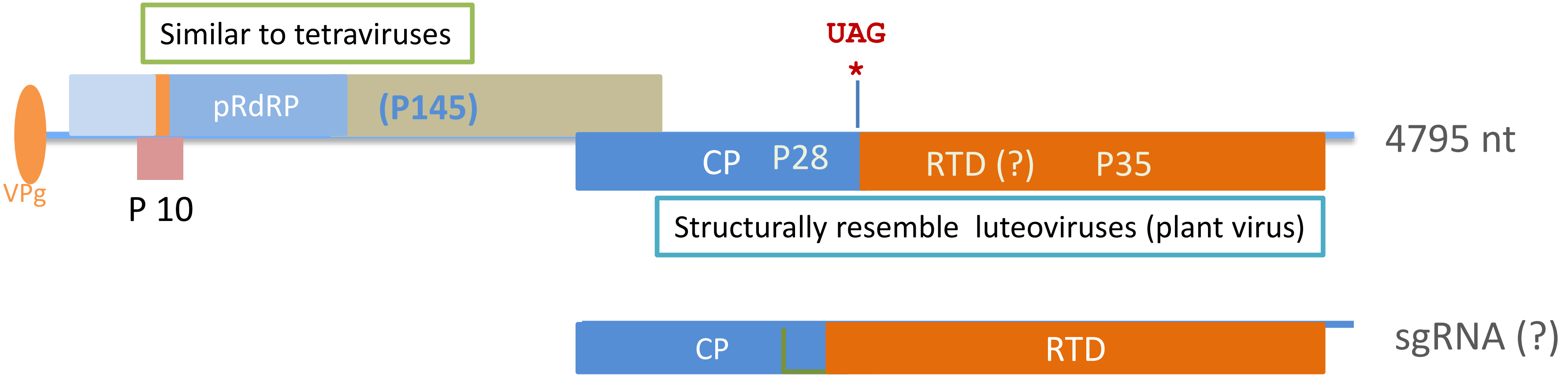
Clone genomic DNA/RNA
- sequence (Sanger sequencing)
- assemble viral genome

Advantages of NGS for Virus Discovery

- Many viruses are latent or asymptomatic
- NGS can identify viral sequences without background information on viruses
- Viral genomes are assembled *de novo* without reference sequences
- NGS has revolutionized virus discovery

Aphis glycines virus (AGV)

-assembled from transcriptome



A new insect virus with tetravirus-like RdRp, and plant virus-like capsid protein

Outline

- Introduction: Why use NGS?
  - Traditional approach for virus discovery
  - Next Generation Sequencing (NGS)
  - Advantages of NGS for virus discovery
- How it’s done
  - Sample selection
  - Sequencing library preparation
  - Sequencing method
  - Assembly of sequencing reads
  - Identification of viral sequence
  - Assembly of viral genome

Sample Selection

- Small sample size (10 ug or less RNA adequate)

  -but the more the better

- Tissue vs. whole organism

  -sequencing depth

- Virus purification

  -helps to identify full-length sequence

  -better approach for DNA viruses

Sequencing Technologies

- Short reads (35-250 nt)
  1. Genome Analyzer IIx (GAIIx), HiSeq2000, HiSeq2500, MiSeq – Illumina

  (Hiseq2000: capable of up to 600Gb per run)

  1. SOLiD 5500xl System – Applied Biosystems
  2. HeliScope™ Single Molecule Sequencer - Helicos

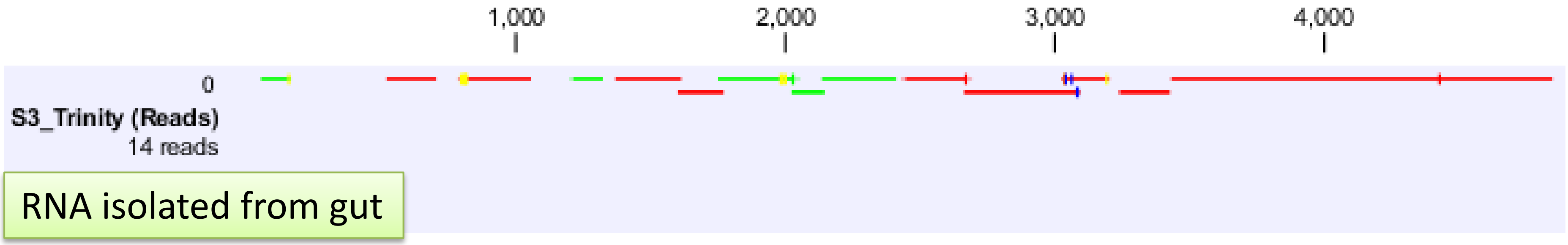
- Long reads (400-20,000 nt)
  1. Genome Sequencer FLX System (454) – Roche
  2. PacBio RS - Pacific Bioscience
  3. Personal Genome Machine, Ion Proton - Ion Torrent
  4. GridION – Oxford Nanopore

Preparation of Sequencing Library

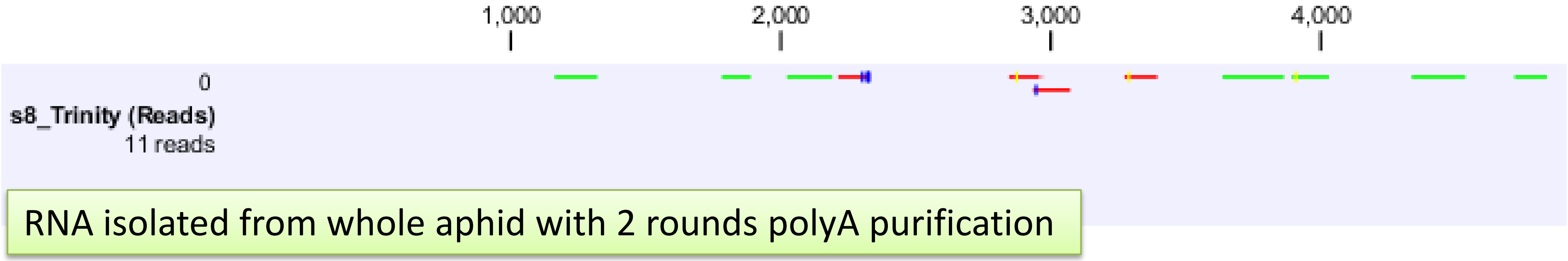
Library type	Viral genomes	Sequence recovery
mRNA*	DNA/RNA	+++, possible full-length
Small RNA	DNA/RNA	+/++
DNA	DNA	+++, possible full-length
DNA or RNA isolated from viruses	DNA/RNA	+++++, full-length

*mRNA purification may result in loss of sequences for viruses that lack polyA tails

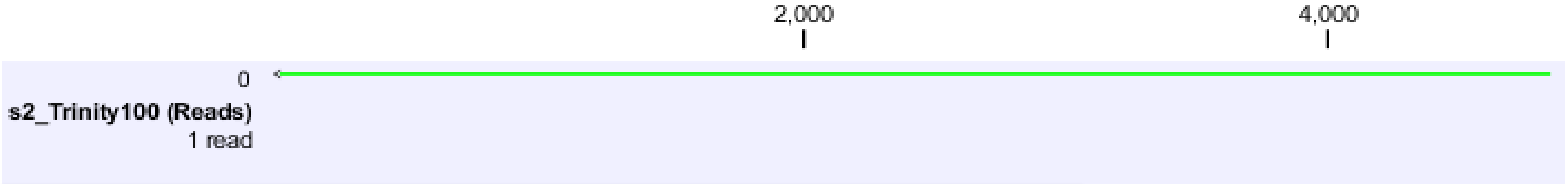
AGV assembled from different sequencing samples



RNA isolated from gut



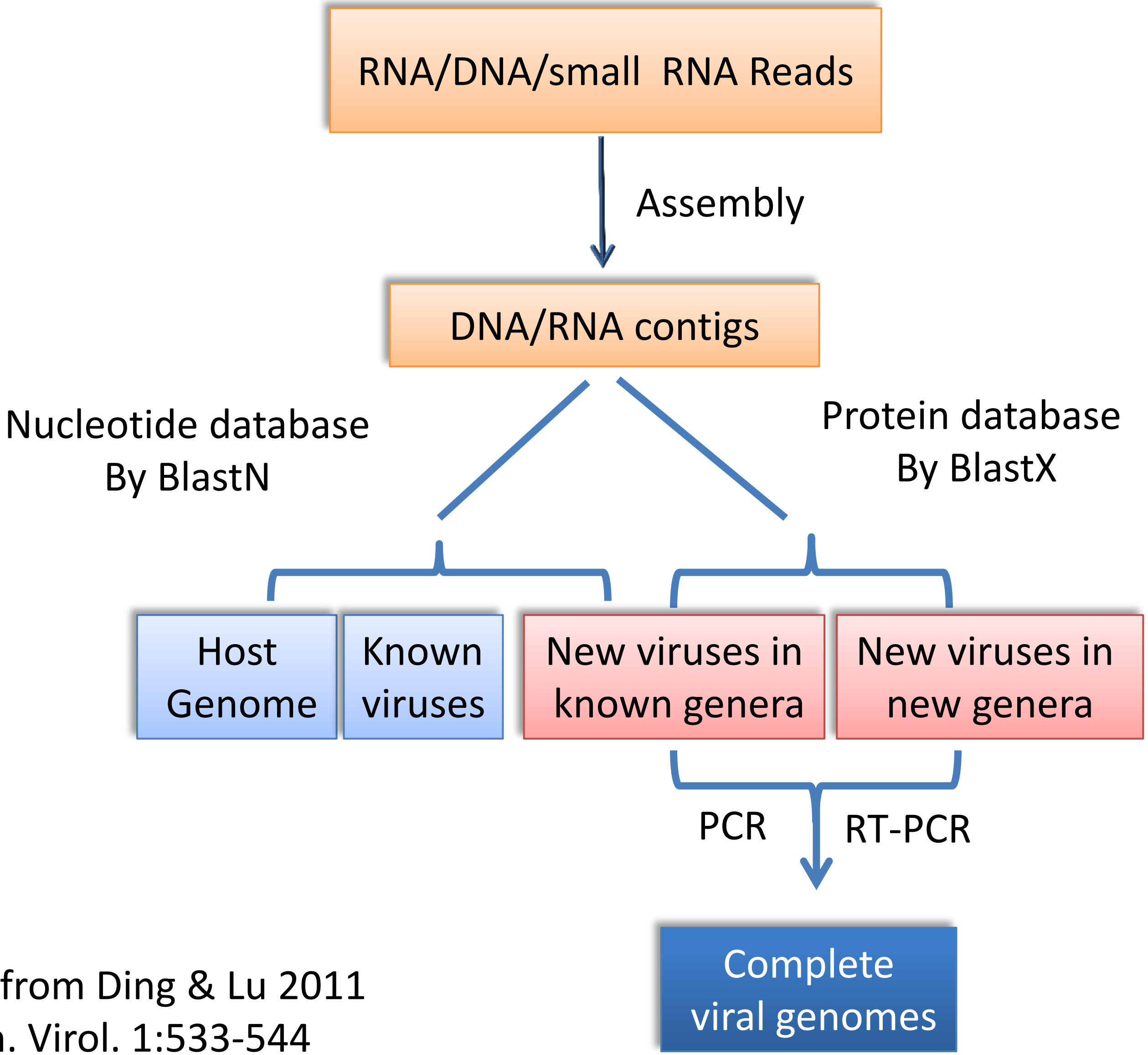
RNA isolated from whole aphid with 2 rounds polyA purification



RNA isolated from whole aphid with 1 round polyA purification

Green: + strand; Red: - strand

NGS for Virus Discovery



RNA/DNA/small RNA Reads

Assembly

DNA/RNA contigs

Nucleotide database
By BlastN

Protein database
By BlastX

Host Genome

Known viruses

New viruses in known genera

New viruses in new genera

PCR

RT-PCR

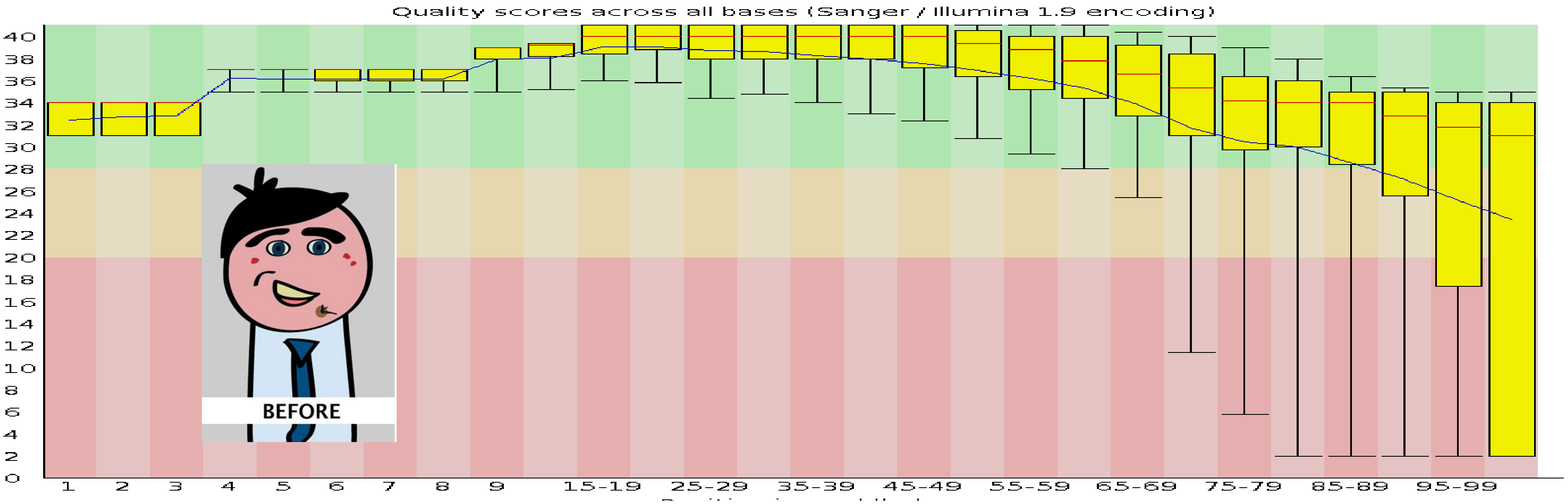
Complete viral genomes

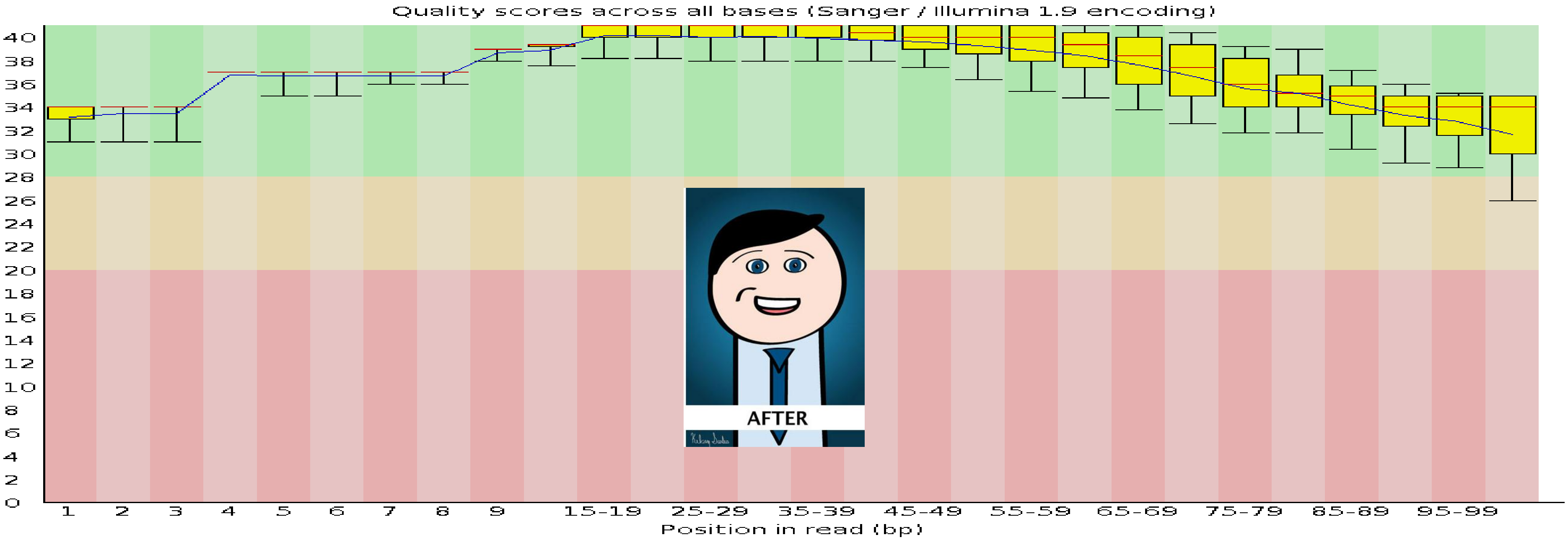
Modified from Ding & Lu 2011
Curr. Opin. Virol. 1:533-544

Assembly of Sequencing Reads
-pre-processing of sequence data

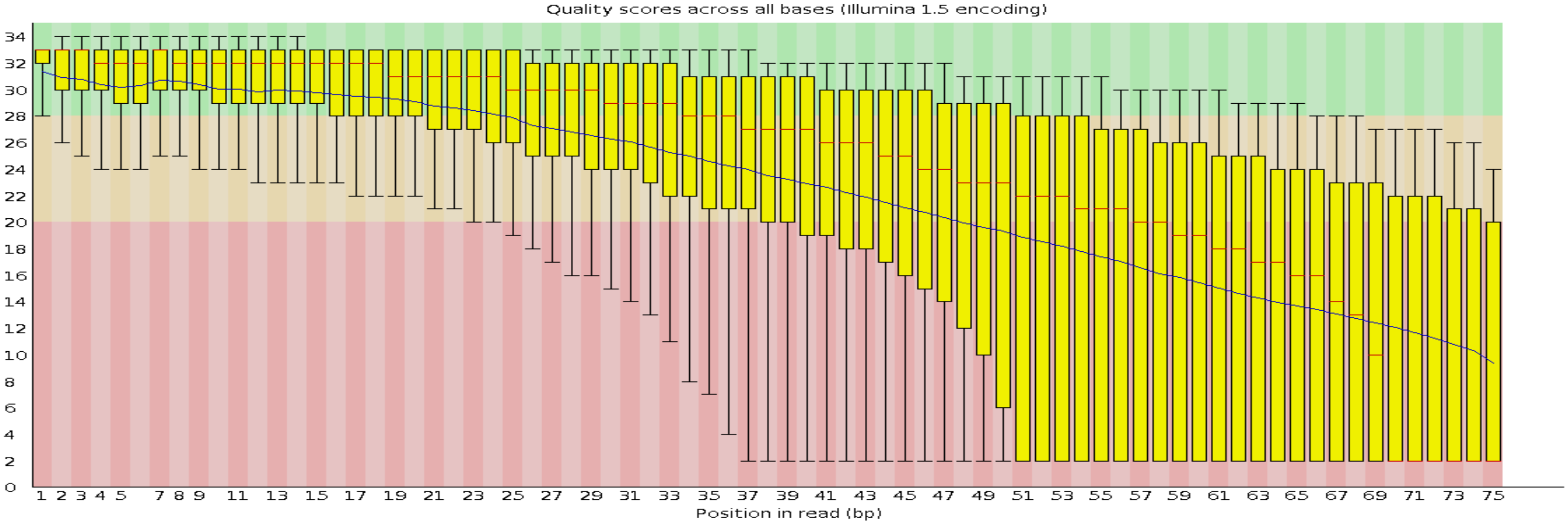
- Remove potential adaptor / index sequences
- Check sequencing quality
  - Quality score; GC content
  - Read length distribution
  - Overrepresented sequences
  - etc.
- If necessary trim bases with low quality

Trimming of Bases with Low QS



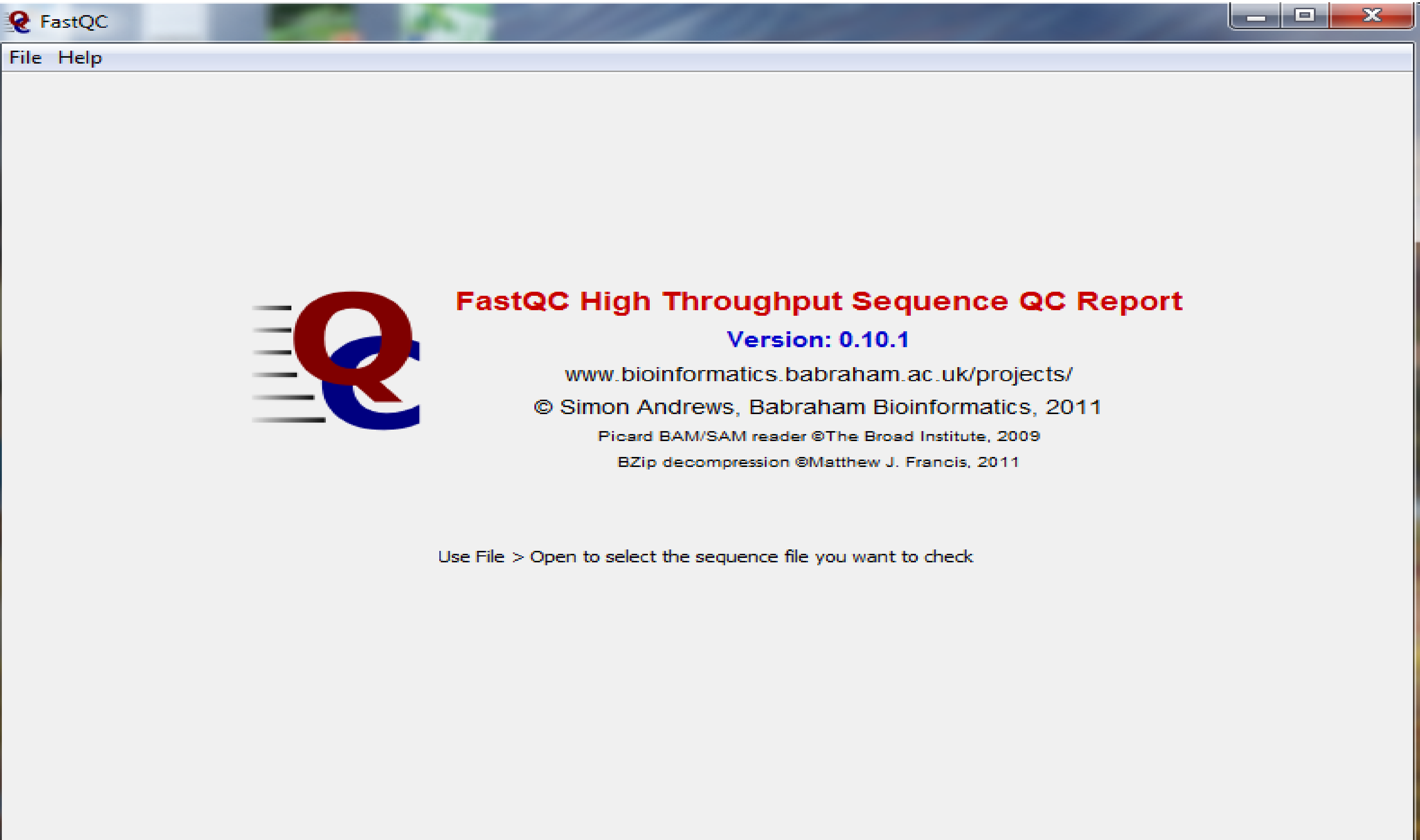


Trimming of bases with low quality scores may result in loss of viral sequences



NOTE: The near full-length genome of AGV was assembled from an untrimmed data set with poor quality scores. The genome could not be assembled from the data set following standard trimming.

Software for Checking Sequence Quality-FastQC



FastQC High Throughput Sequence QC Report

Version: 0.10.1

www.bioinformatics.babraham.ac.uk/projects/

© Simon Andrews, Babraham Bioinformatics, 2011

Picard BAM/SAM reader ©The Broad Institute, 2009

BZip decompression ©Matthew J. Francis, 2011

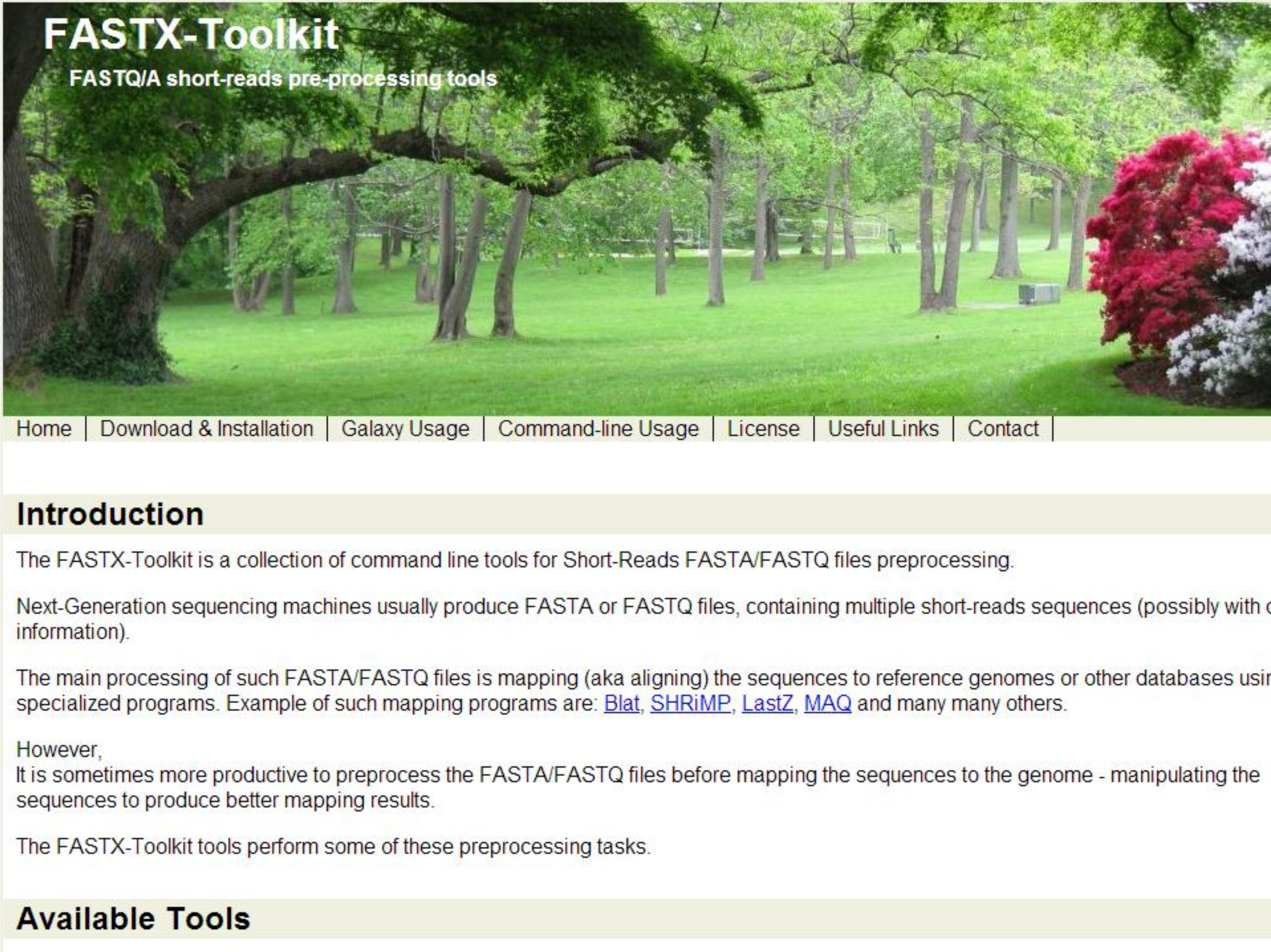
Use File > Open to select the sequence file you want to check

Overrepresented Sequences

Sequence	Count	Percentage	Possible Source
AGATCGGAAGAGCACACGTCTGAACTCCAGTCACCTTGTAATCTCGTATG	1968861	2.20	TruSeq Adapter, Index 12 (100% over 49bp)

Sequence	Count	Percentage	Possible Source
CAGATTTCGGGCTAAAGGGAATACGGTTAAAATCCCGTGACCTGCCCTGT	51018488	40.90	No Hit
TCAGATTTCGGGCTAAAGGGAATACGGTTAAAATCCCGTGACCTGCCCTG	24264170	19.45	No Hit

The seqeunces were derived from *Penaeus vannamei* 18S ribosomal RNA -cotaminated in sRNA



FASTX-Toolkit

FASTQ/A short-reads pre-processing tools

Introduction

The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

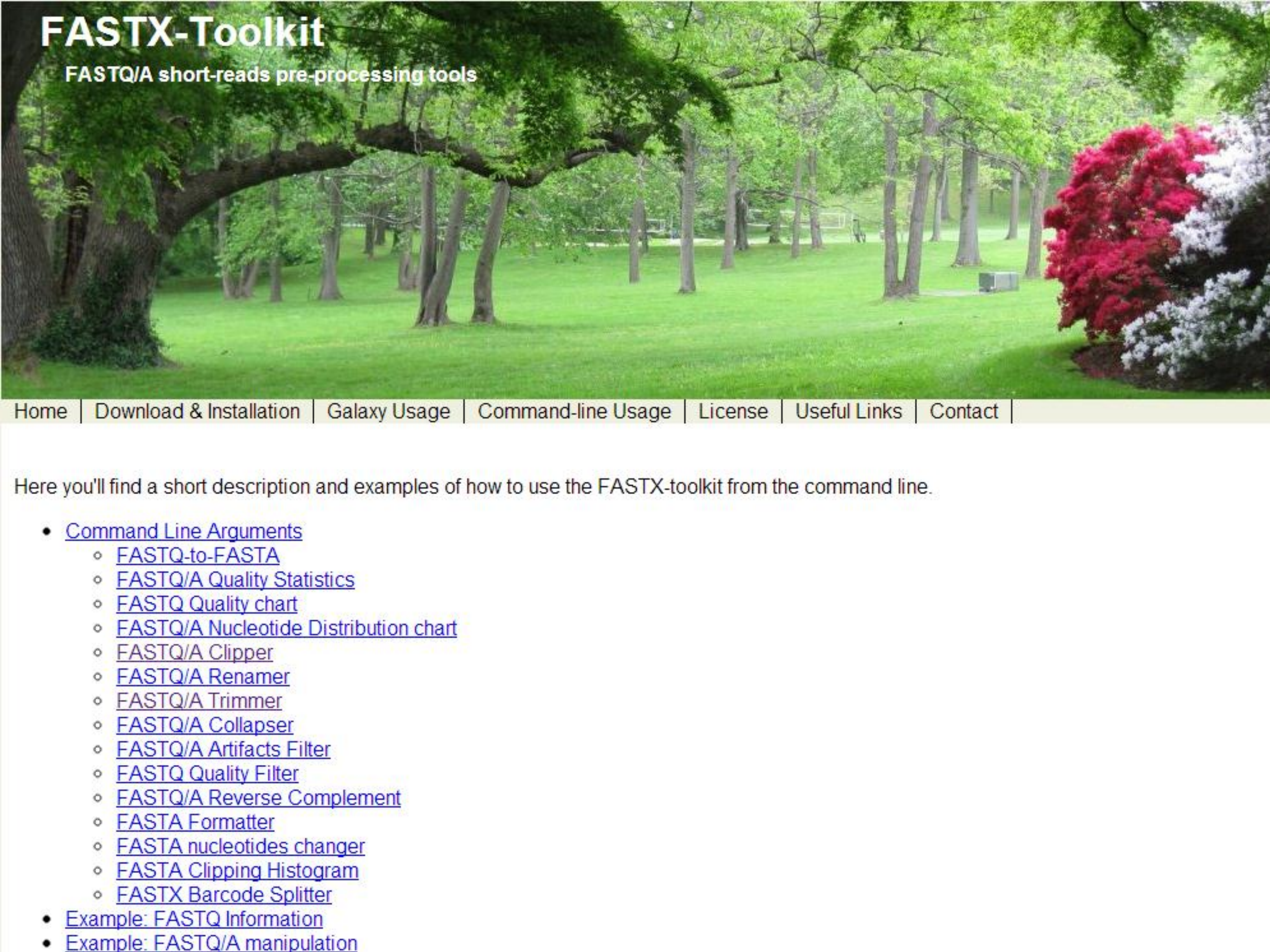
Next-Generation sequencing machines usually produce FASTA or FASTQ files, containing multiple short-reads sequences (possibly with c
information).

The main processing of such FASTA/FASTQ files is mapping (aka aligning) the sequences to reference genomes or other databases usir
specialized programs. Example of such mapping programs are: Blat, SHRiMP, LastZ, MAQ and many many others.

However,
It is sometimes more productive to preprocess the FASTA/FASTQ files before mapping the sequences to the genome - manipulating the sequences to produce better mapping results.

The FASTX-Toolkit tools perform some of these preprocessing tasks.

Available Tools

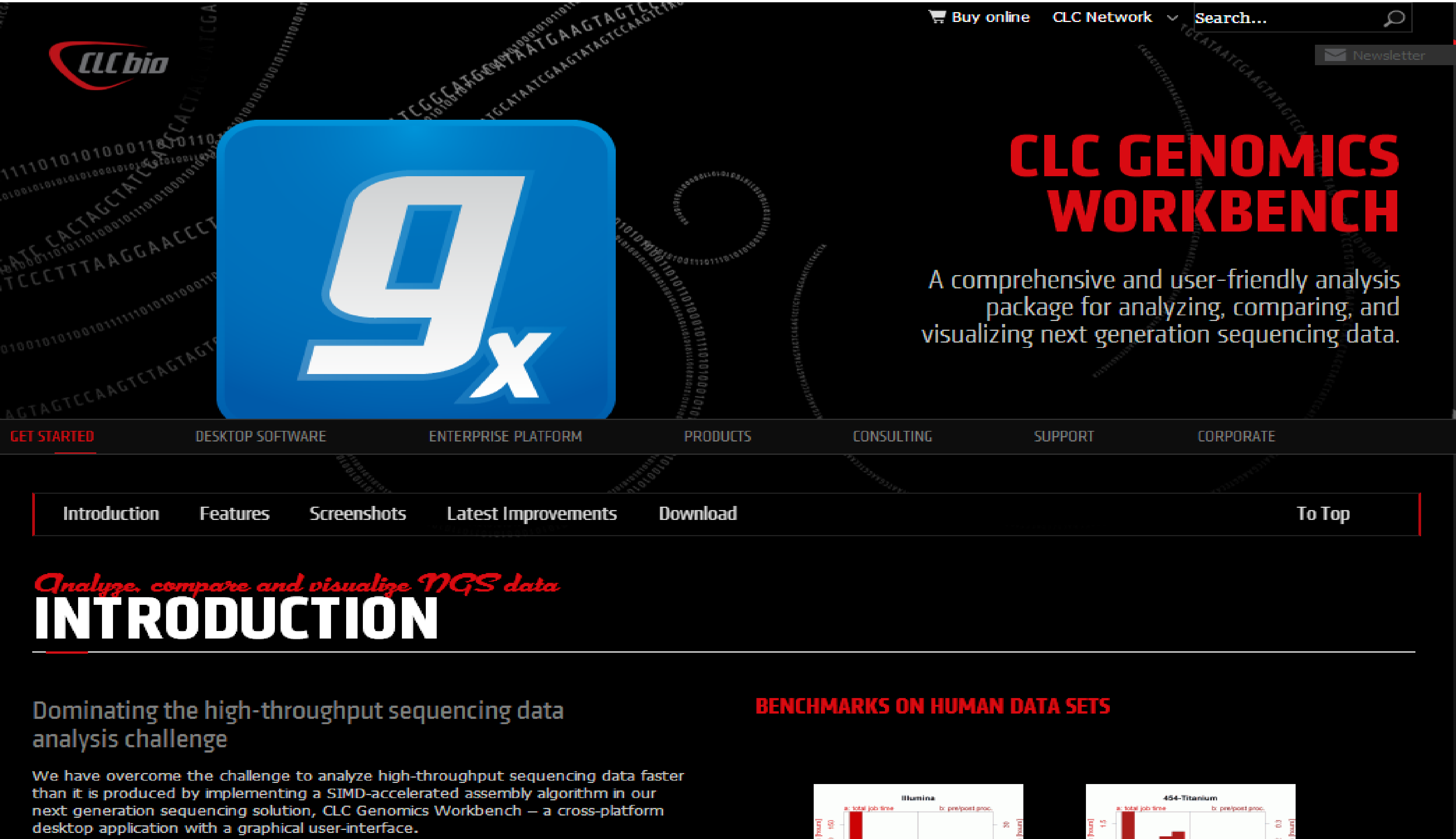

FASTX-Toolkit
FASTQ/A short-reads pre-processing tools
Home | Download & Installation | Galaxy Usage | Command-line Usage | License | Useful Links | Contact

Here you'll find a short description and examples of how to use the FASTX-toolkit from the command line.

- Command Line Arguments
  - FASTQ-to-FASTA
  - FASTQ/A Quality Statistics
  - FASTQ Quality chart
  - FASTQ/A Nucleotide Distribution chart
  - FASTQ/A Clipper
  - FASTQ/A Renamer
  - FASTQ/A Trimmer
  - FASTQ/A Collapser
  - FASTQ/A Artifacts Filter
  - FASTQ Quality Filter
  - FASTQ/A Reverse Complement
  - FASTA Formatter
  - FASTA nucleotides changer
  - FASTA Clipping Histogram
  - FASTX Barcode Splitter
- Example: FASTQ Information
- Example: FASTQ/A manipulation

CLC Genomics Workbench

(US$5,000 per copy , >US$1,000/per year for update)



Buy online CLC Network Search...

Newsletter

CLC bio

CLC GENOMICS WORKBENCH

A comprehensive and user-friendly analysis package for analyzing, comparing, and visualizing next generation sequencing data.

GET STARTED DESKTOP SOFTWARE ENTERPRISE PLATFORM PRODUCTS CONSULTING SUPPORT CORPORATE

Introduction Features Screenshots Latest Improvements Download To Top

Analyze, compare and visualize NGS data

INTRODUCTION

Dominating the high-throughput sequencing data analysis challenge

We have overcome the challenge to analyze high-throughput sequencing data faster than it is produced by implementing a SIMD-accelerated assembly algorithm in our next generation sequencing solution, CLC Genomics Workbench – a cross-platform desktop application with a graphical user-interface.

BENCHMARKS ON HUMAN DATA SETS

Assembly of Sequencing Reads

- *de novo* assembly or mapping (alignment)

  -*de novo* assembly: searching for new viruses, no reference is needed

  -mapping: re-sequencing, SNP, isolate, need reference sequences (MARA, GATK and other toolkits)

- *de novo* assembly may provide extra information about known viral sequences

  Shrimp virus: Infectious myonecrosis virus (IMNV, a dsRNA virus)

  - documented seq: 7560 bp (Poulos et al., JGV, 2006 87: 987-996)

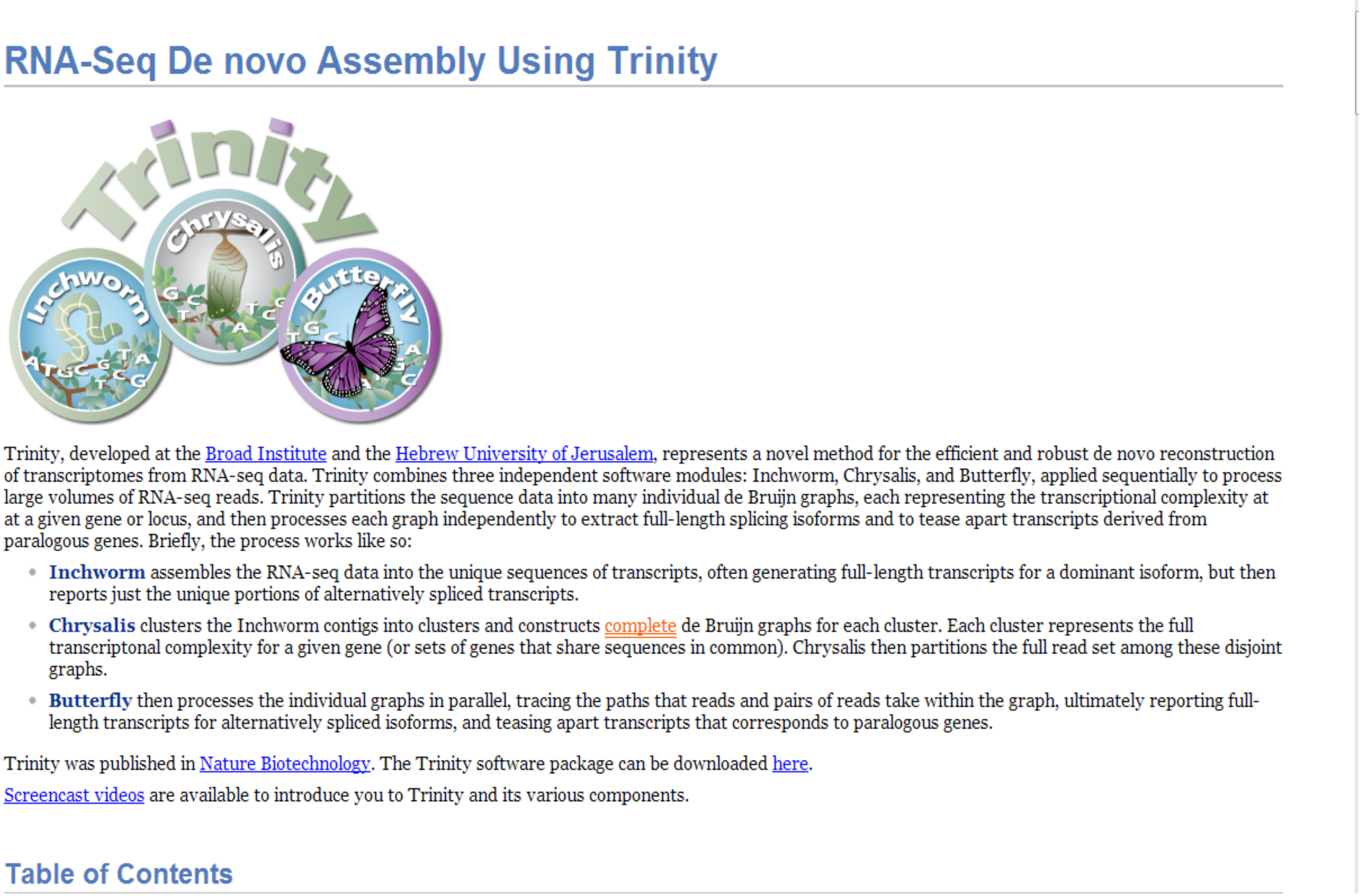
  - *de novo* assembled from RNA-seq: 8233 bp

  RT-PCR proved IMNV should have at least 8233 bp

Thursday 9:45 am 168 Virus 4 ; Duan Loy

BROAD INSTITUTE

Trinity for Assembly



RNA-Seq De novo Assembly Using Trinity

Trinity, developed at the Broad Institute and the Hebrew University of Jerusalem, represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-seq data. Trinity combines three independent software modules: Inchworm, Chrysalis, and Butterfly, applied sequentially to process large volumes of RNA-seq reads. Trinity partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes. Briefly, the process works like so:

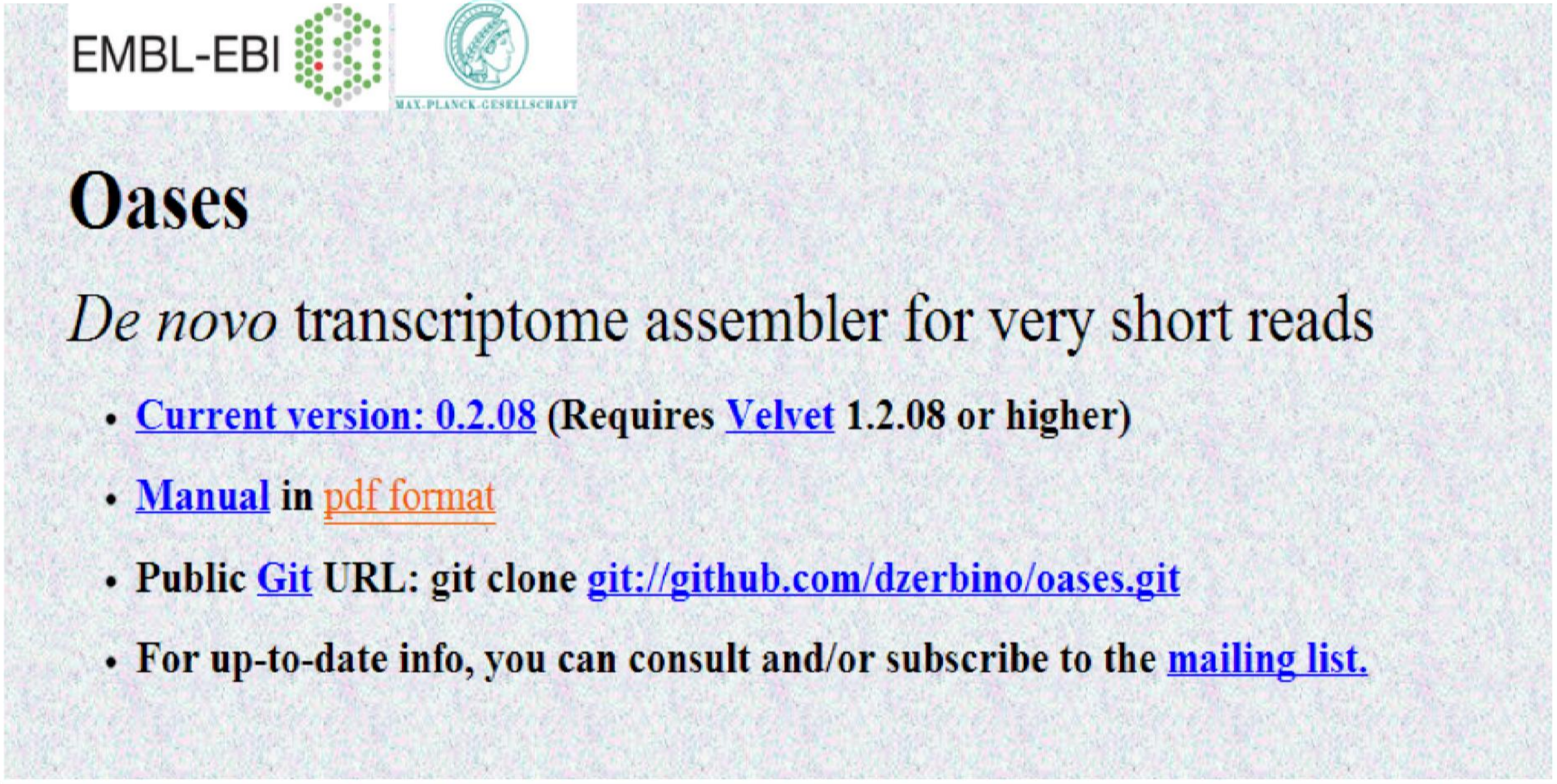
- **Inchworm** assembles the RNA-seq data into the unique sequences of transcripts, often generating full-length transcripts for a dominant isoform, but then reports just the unique portions of alternatively spliced transcripts.
- **Chrysalis** clusters the Inchworm contigs into clusters and constructs complete de Bruijn graphs for each cluster. Each cluster represents the full transcriptonal complexity for a given gene (or sets of genes that share sequences in common). Chrysalis then partitions the full read set among these disjoint graphs.
- **Butterfly** then processes the individual graphs in parallel, tracing the paths that reads and pairs of reads take within the graph, ultimately reporting full-length transcripts for alternatively spliced isoforms, and teasing apart transcripts that corresponds to paralogous genes.

Trinity was published in Nature Biotechnology. The Trinity software package can be downloaded here.

Screencast videos are available to introduce you to Trinity and its various components.

Table of Contents

Oases/Velvet for Assembly



Oases

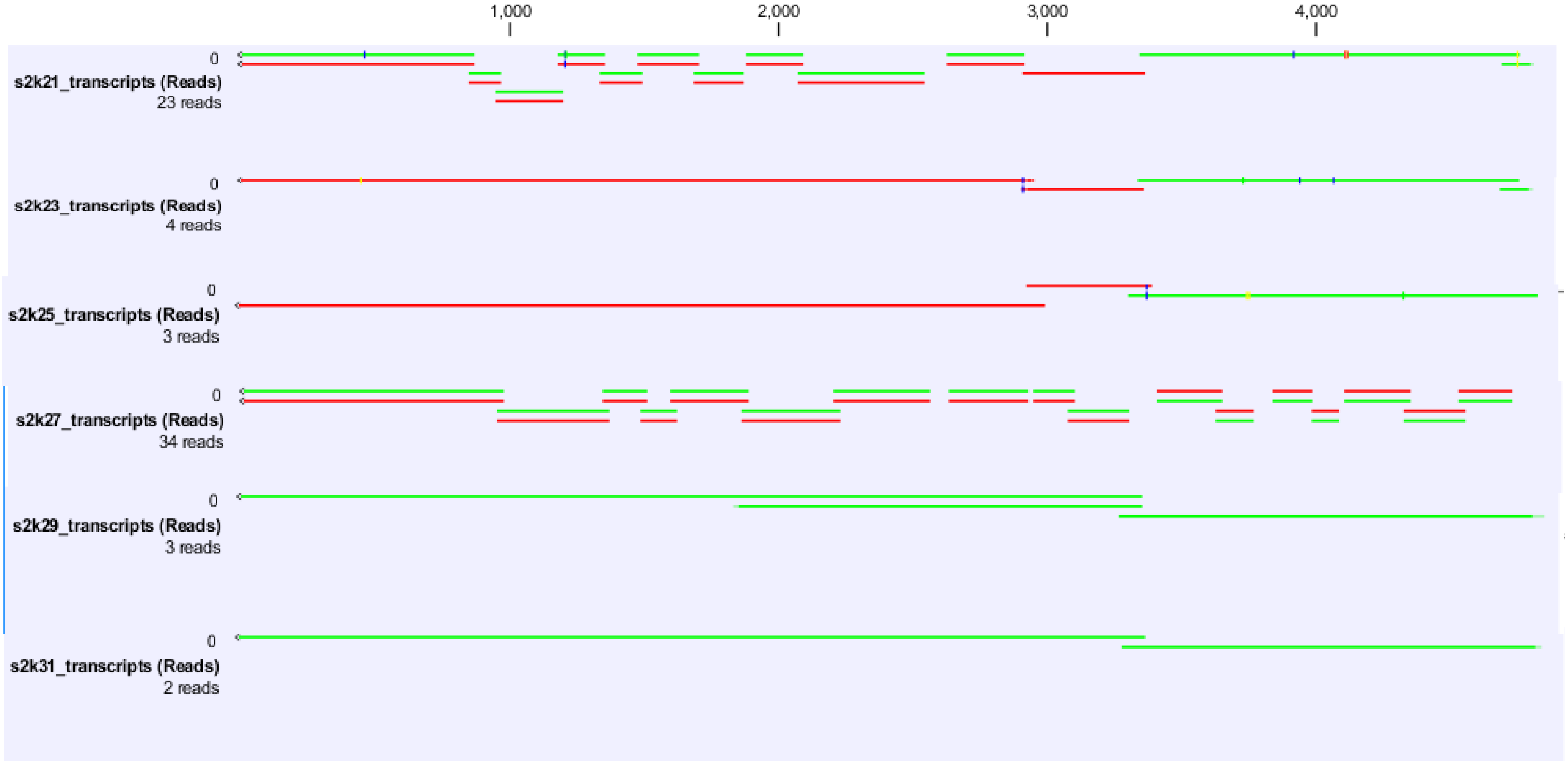
De novo transcriptome assembler for very short reads

- Current version: 0.2.08 (Requires Velvet 1.2.08 or higher)
- Manual in pdf format
- Public Git URL: git clone git://github.com/dzerbino/oases.git
- For up-to-date info, you can consult and/or subscribe to the mailing list.

Running the Assembly Program

- Two most important parameters for assembly
  - K-mers (word length): length of sequence fragments used for joining
  - C - coverage cut-off
- Different combinations of K and C will result in assembly of different contigs
- Multiple K and C should be tested for best results (Liu et al. PLoS One. 2012;7(9):e45161. doi: 10.1371/journal.pone)

Multiple K Test for Assembly of AGV using Oases/Velvet



(Here) read = contig

Green: + strand; Red: - strand

Data Analysis

How do we find viral sequences?

- Annotation of contigs

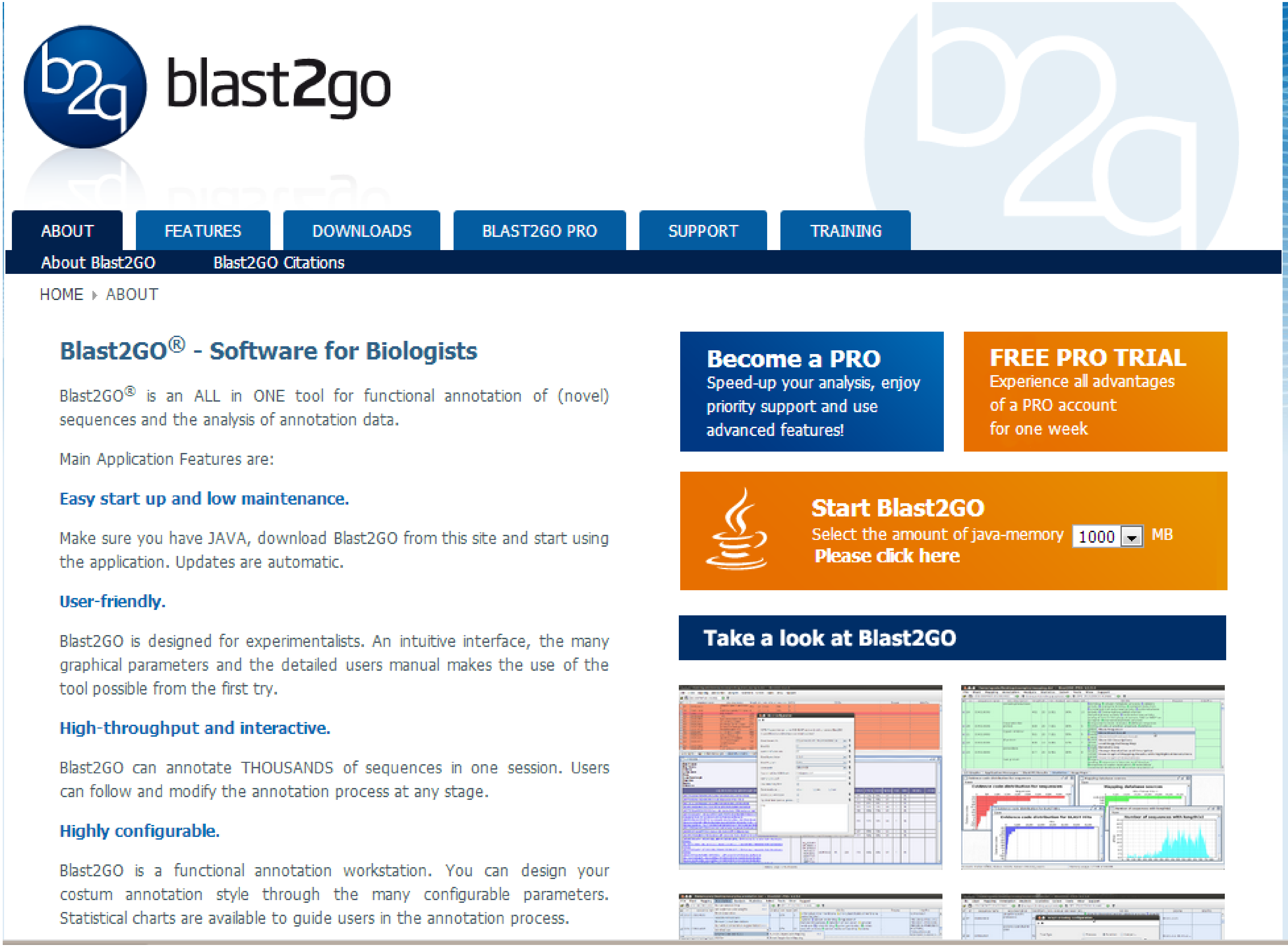
    -search for viral genes using BLASTx or BLASTn

- BLAST against NCBI database
- BLAST using your own databases
- Blast2GO platform

    -annotation of contigs

    -motif search

    -analysis of annotation data



Blast2GO® - Software for Biologists

Blast2GO® is an ALL in ONE tool for functional annotation of (novel) sequences and the analysis of annotation data.

Main Application Features are:

Easy start up and low maintenance.

Make sure you have JAVA, download Blast2GO from this site and start using the application. Updates are automatic.

User-friendly.

Blast2GO is designed for experimentalists. An intuitive interface, the many graphical parameters and the detailed users manual makes the use of the tool possible from the first try.

High-throughput and interactive.

Blast2GO can annotate THOUSANDS of sequences in one session. Users can follow and modify the annotation process at any stage.

Highly configurable.

Blast2GO is a functional annotation workstation. You can design your costum annotation style through the many configurable parameters. Statistical charts are available to guide users in the annotation process.

Become a PRO
Speed-up your analysis, enjoy priority support and use advanced features!

FREE PRO TRIAL
Experience all advantages of a PRO account for one week

Start Blast2GO
Select the amount of java-memory 1000 ▾ MB
Please click here

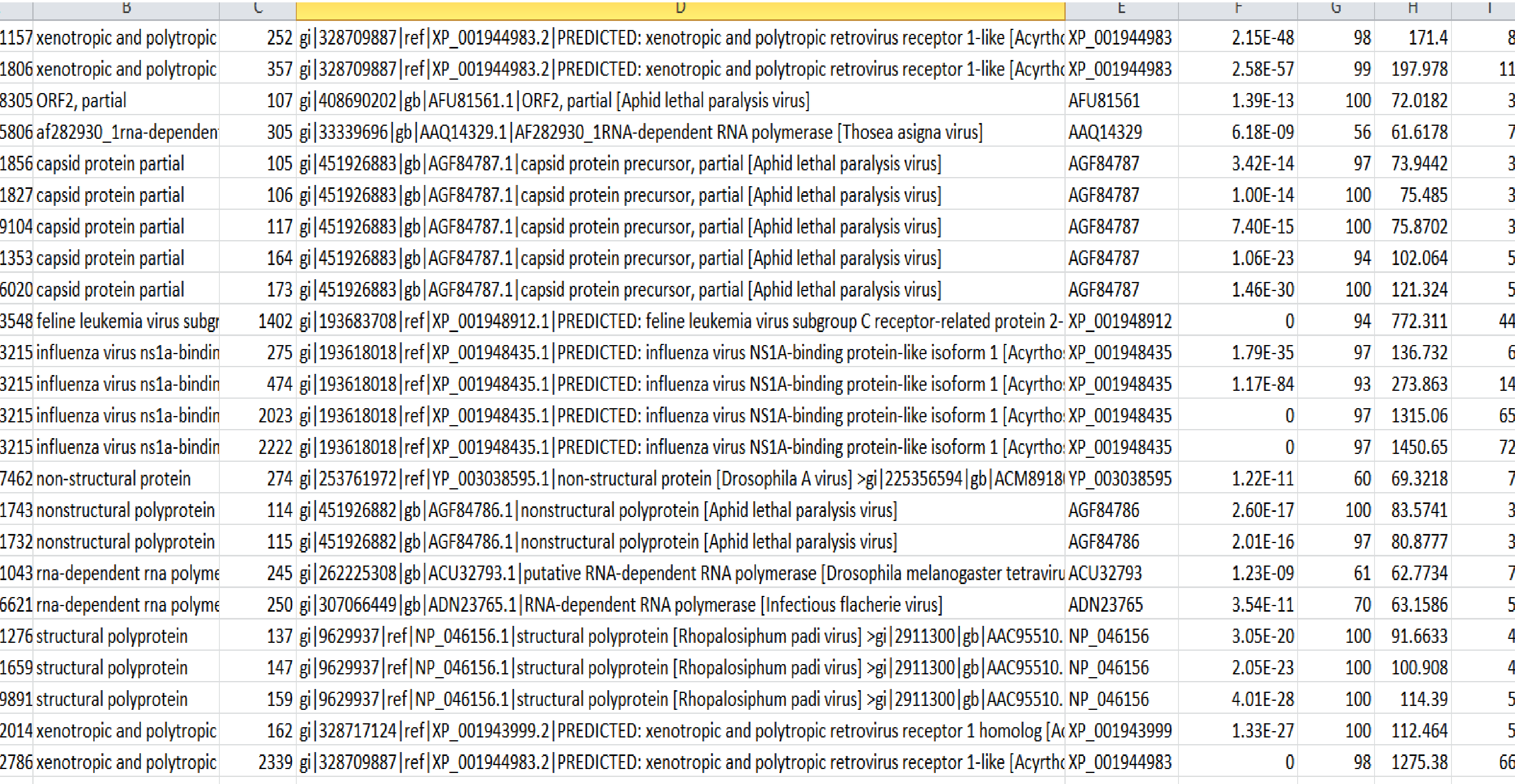
Take a look at Blast2GO

Data Analysis

Analyzing virus-derived contigs

- Extract BLAST data (sequences with virus as top hit)
- Organize contigs that hit the same or similar viruses
- Join contigs into viral genome
- Design primers for PCR/RT-PCR to fill sequence gaps
- Sequence to confirm *in silico* cloning result
- 5’ and 3’ RACE to identify end sequences

Working with Viral Contigs

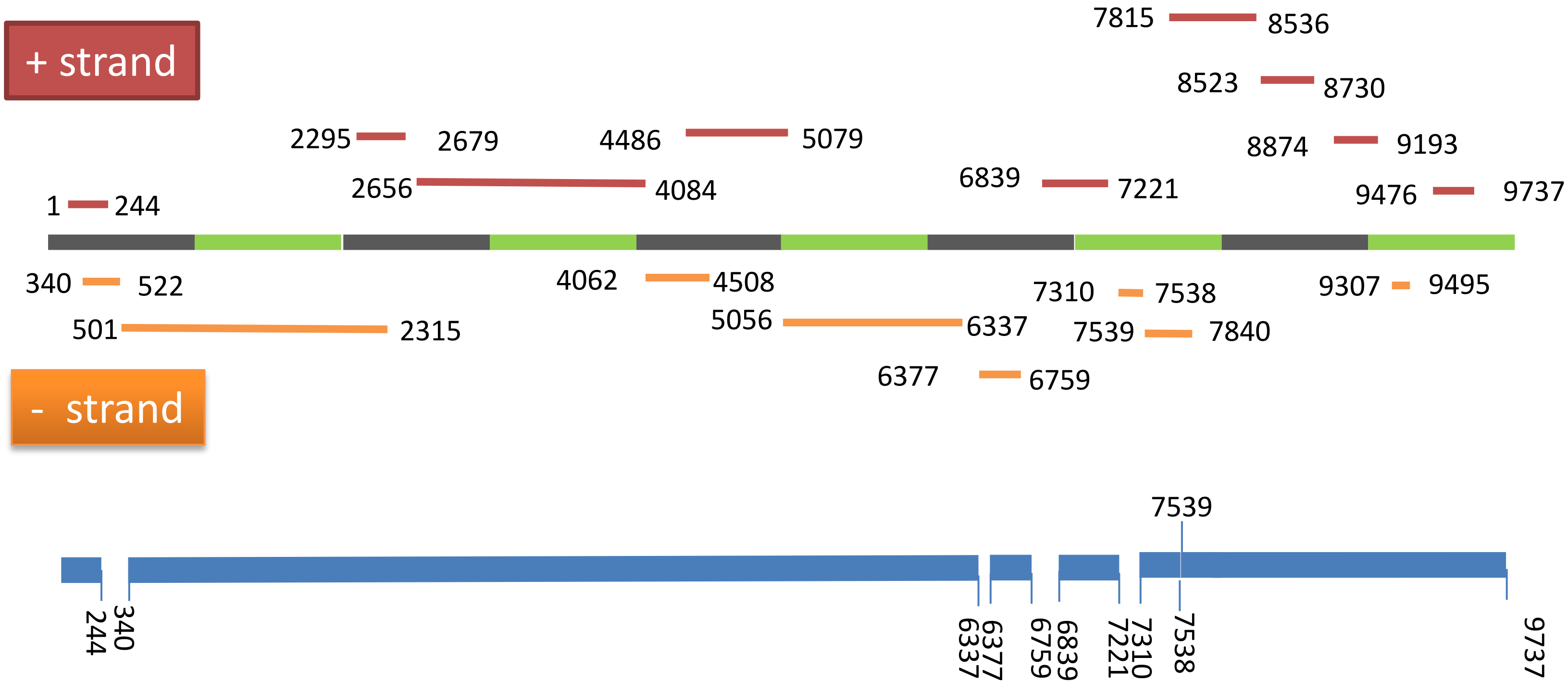


	B	C	D	E	F	G	H	I
1157	xenotropic and polytropic	252	gi|328709887|ref|XP_001944983.2|PREDICTED: xenotropic and polytropic retrovirus receptor 1-like [Acyrthc	XP_001944983	2.15E-48	98	171.4	8
1806	xenotropic and polytropic	357	gi|328709887|ref|XP_001944983.2|PREDICTED: xenotropic and polytropic retrovirus receptor 1-like [Acyrthc	XP_001944983	2.58E-57	99	197.978	11
8305	ORF2, partial	107	gi|408690202|gb|AFU81561.1|ORF2, partial [Aphid lethal paralysis virus]	AFU81561	1.39E-13	100	72.0182	3
5806	af282930_1rna-dependen	305	gi|33339696|gb|AAQ14329.1|AF282930_1RNA-dependent RNA polymerase [Thosea asigna virus]	AAQ14329	6.18E-09	56	61.6178	7
1856	capsid protein partial	105	gi|451926883|gb|AGF84787.1|capsid protein precursor, partial [Aphid lethal paralysis virus]	AGF84787	3.42E-14	97	73.9442	3
1827	capsid protein partial	106	gi|451926883|gb|AGF84787.1|capsid protein precursor, partial [Aphid lethal paralysis virus]	AGF84787	1.00E-14	100	75.485	3
9104	capsid protein partial	117	gi|451926883|gb|AGF84787.1|capsid protein precursor, partial [Aphid lethal paralysis virus]	AGF84787	7.40E-15	100	75.8702	3
1353	capsid protein partial	164	gi|451926883|gb|AGF84787.1|capsid protein precursor, partial [Aphid lethal paralysis virus]	AGF84787	1.06E-23	94	102.064	5
6020	capsid protein partial	173	gi|451926883|gb|AGF84787.1|capsid protein precursor, partial [Aphid lethal paralysis virus]	AGF84787	1.46E-30	100	121.324	5
3548	feline leukemia virus subgr	1402	gi|193683708|ref|XP_001948912.1|PREDICTED: feline leukemia virus subgroup C receptor-related protein 2-	XP_001948912	0	94	772.311	44
3215	influenza virus ns1a-bindin	275	gi|193618018|ref|XP_001948435.1|PREDICTED: influenza virus NS1A-binding protein-like isoform 1 [Acyrtho:	XP_001948435	1.79E-35	97	136.732	6
3215	influenza virus ns1a-bindin	474	gi|193618018|ref|XP_001948435.1|PREDICTED: influenza virus NS1A-binding protein-like isoform 1 [Acyrtho:	XP_001948435	1.17E-84	93	273.863	14
3215	influenza virus ns1a-bindin	2023	gi|193618018|ref|XP_001948435.1|PREDICTED: influenza virus NS1A-binding protein-like isoform 1 [Acyrtho:	XP_001948435	0	97	1315.06	65
3215	influenza virus ns1a-bindin	2222	gi|193618018|ref|XP_001948435.1|PREDICTED: influenza virus NS1A-binding protein-like isoform 1 [Acyrtho:	XP_001948435	0	97	1450.65	72
7462	non-structural protein	274	gi|253761972|ref|YP_003038595.1|non-structural protein [Drosophila A virus] >gi|225356594|gb|ACM8918	YP_003038595	1.22E-11	60	69.3218	7
1743	nonstructural polyprotein	114	gi|451926882|gb|AGF84786.1|nonstructural polyprotein [Aphid lethal paralysis virus]	AGF84786	2.60E-17	100	83.5741	3
1732	nonstructural polyprotein	115	gi|451926882|gb|AGF84786.1|nonstructural polyprotein [Aphid lethal paralysis virus]	AGF84786	2.01E-16	97	80.8777	3
1043	rna-dependent rna polyme	245	gi|262225308|gb|ACU32793.1|putative RNA-dependent RNA polymerase [Drosophila melanogaster tetraviru	ACU32793	1.23E-09	61	62.7734	7
6621	rna-dependent rna polyme	250	gi|307066449|gb|ADN23765.1|RNA-dependent RNA polymerase [Infectious flacherie virus]	ADN23765	3.54E-11	70	63.1586	5
1276	structural polyprotein	137	gi|9629937|ref|NP_046156.1|structural polyprotein [Rhopalosiphum padi virus] >gi|2911300|gb|AAC95510.	NP_046156	3.05E-20	100	91.6633	4
1659	structural polyprotein	147	gi|9629937|ref|NP_046156.1|structural polyprotein [Rhopalosiphum padi virus] >gi|2911300|gb|AAC95510.	NP_046156	2.05E-23	100	100.908	4
9891	structural polyprotein	159	gi|9629937|ref|NP_046156.1|structural polyprotein [Rhopalosiphum padi virus] >gi|2911300|gb|AAC95510.	NP_046156	4.01E-28	100	114.39	5
2014	xenotropic and polytropic	162	gi|328717124|ref|XP_001943999.2|PREDICTED: xenotropic and polytropic retrovirus receptor 1 homolog [A(	XP_001943999	1.33E-27	100	112.464	5
2786	xenotropic and polytropic	2339	gi|328709887|ref|XP_001944983.2|PREDICTED: xenotropic and polytropic retrovirus receptor 1-like [Acyrthc	XP_001944983	0	98	1275.38	66

viral gene ~~==~~ virus

Trinity Assembly of APV2 (>9800 nt)

Assembled using sRNA isolated from pea aphid

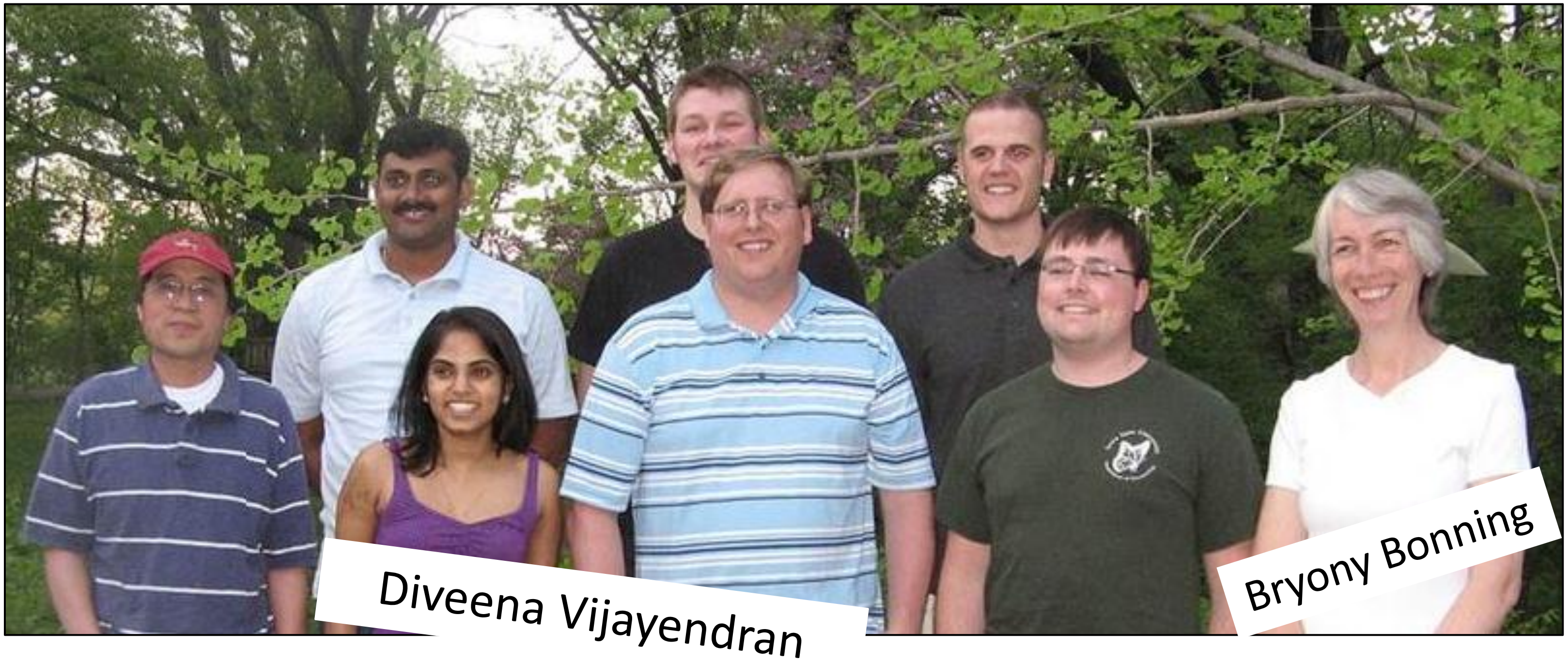


APV2-Acyrthosiphon pisum virus 2 (dicistrovrius)

Summary

- No single rule can be used to find a virus by NGS
- Knowledge of virology can greatly help for analyzing NGS data
- Manual alignment of virus derived sequences may be needed
- Biological evidence is required for verifying true nature of viral sequences discovered by NGS

Acknowledgements



Diveena Vijayendran

Bryony Bonning

John K. VanDyk
Lyric Bartholomay
Duan Loy

IOWA STATE UNIVERSITY
OF SCIENCE AND TECHNOLOGY

CIAG
Center for Integrated Animal Genomics

VII
VIRUS-INSECT INTERACTIONS