Novel Virus Discovery Pipeline

Pipeline : a set of data processing elements connected in series
output of one element is the input of the next one - Wikipedia

Virus Discovery Pipeline : pipeline that is for identifying
known or unknown virus using deep sequencing data

Several pipelines are available online

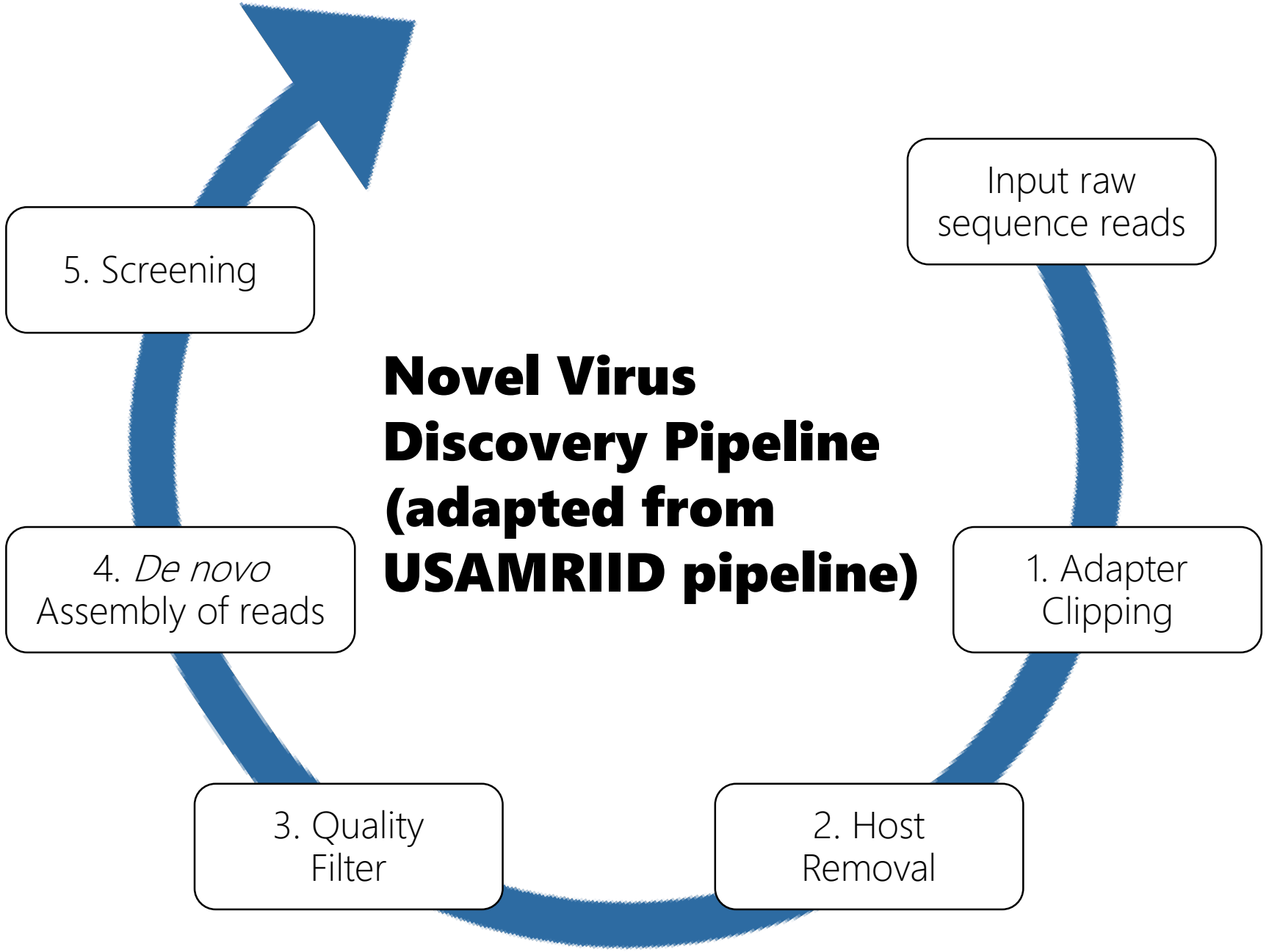
Name	Author	Development organization, Nation	No. of citation
VirFind	Thien Ho, et al.	University of Arkansas System, USA	51
VirusSeeker	Guoyan Zhao, et al.	Washington University School of Medicine, USA	9
VirusDetect	Yi Zheng, et al.	Boyce Thompson Institute, USA	13
VIP	Yang Li, et al.	China CDC, China	14

But, There are some drawbacks

1) Hard to Use : Need to install many of prerequisite programs and understand decades of options

2) Complex to modify : Having complicated dependency structure, too complex to modify and optimize

Need to develop handy and simple Virus discovery pipeline


Input raw sequence reads
1. Adapter Clipping
2. Host Removal
3. Quality Filter
4. *De novo* Assembly of reads
5. Screening
Novel Virus Discovery Pipeline (adapted from USAMRIID pipeline)

1. Adapter Clipping

Objective | Removing adapter sequences which are used for NGS process

Program | Trimmomatic (2014)

"There are slight differences between trimming programs"

- An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis (2013)

AND

Good quality NGS data

⬇

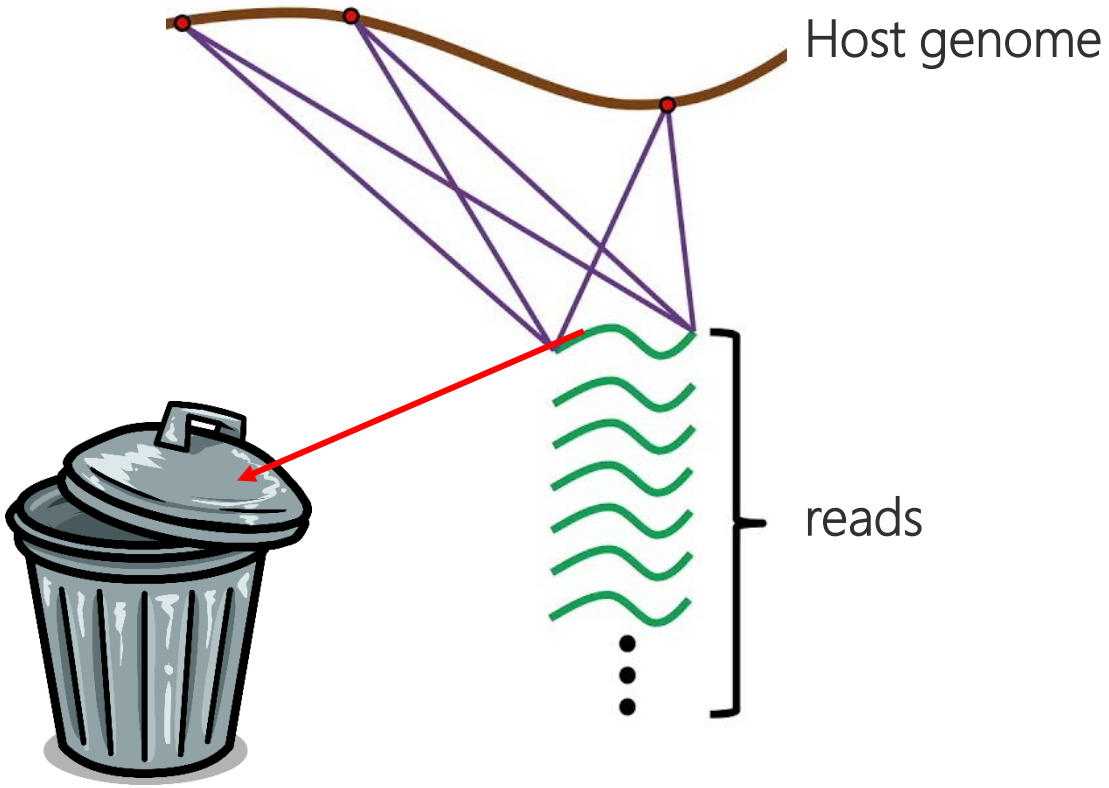
Aggressive and Speedy trimmer

in Truseq preparation kit

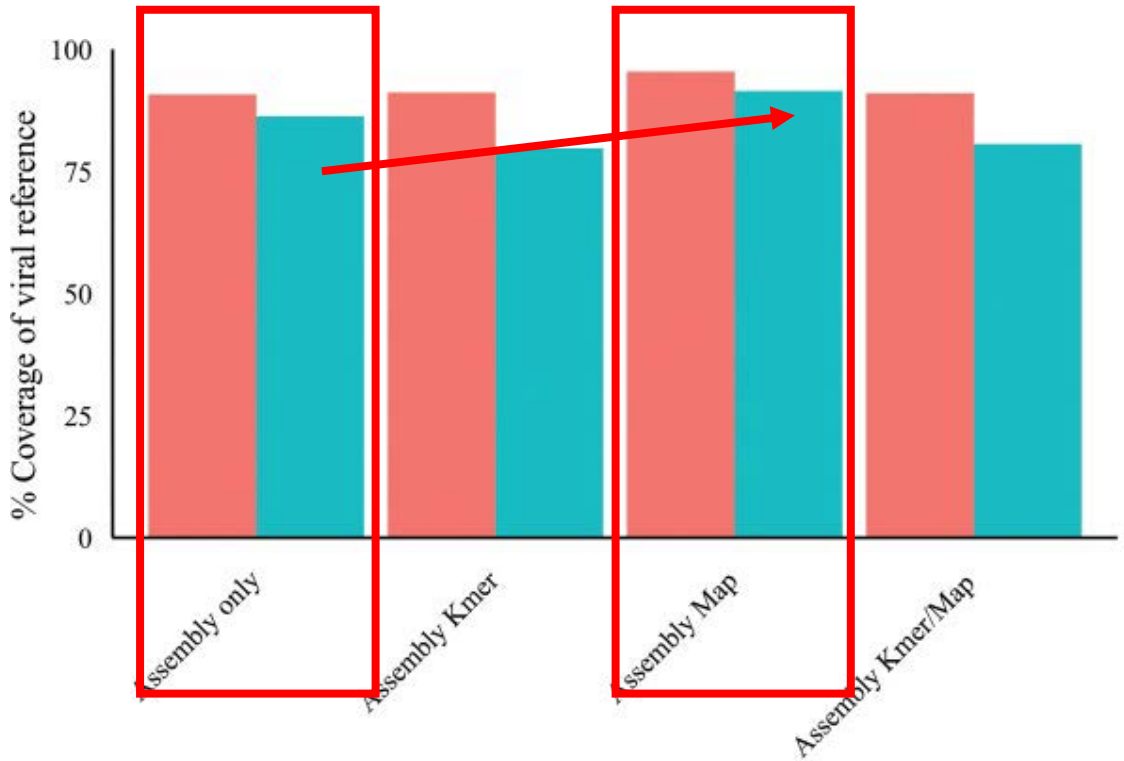
2. Host removal

Objective | Removing host genome sequences

Program | Bowtie2 (2012)



Organism	Source	Note
PhiX	NCBI	PhiX174
Human	NCBI	GRCh38
Mouse_Rat	URSDB	Mus musculus Rattus Norvegicus
Sorex	URSDB	Sorex araneus
Cro	URSDB	Crocidura attenuate
Penguin	URSDB	Pteria penguin
Bat	URSDB	Rhinolophus ferrumequinum
Tick	URSDB	Haemaphysalis longicornis

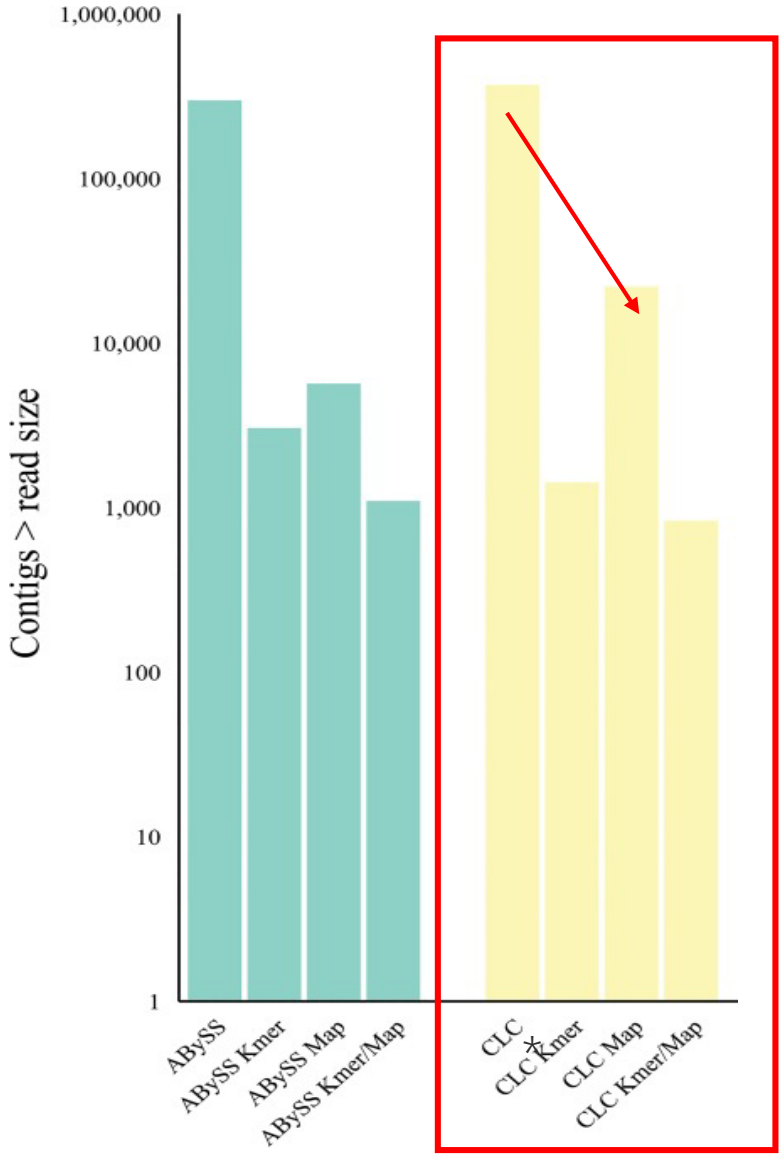


(aligning all contigs made)

Total reference coverage

Largest contig

(aligning only one contig)



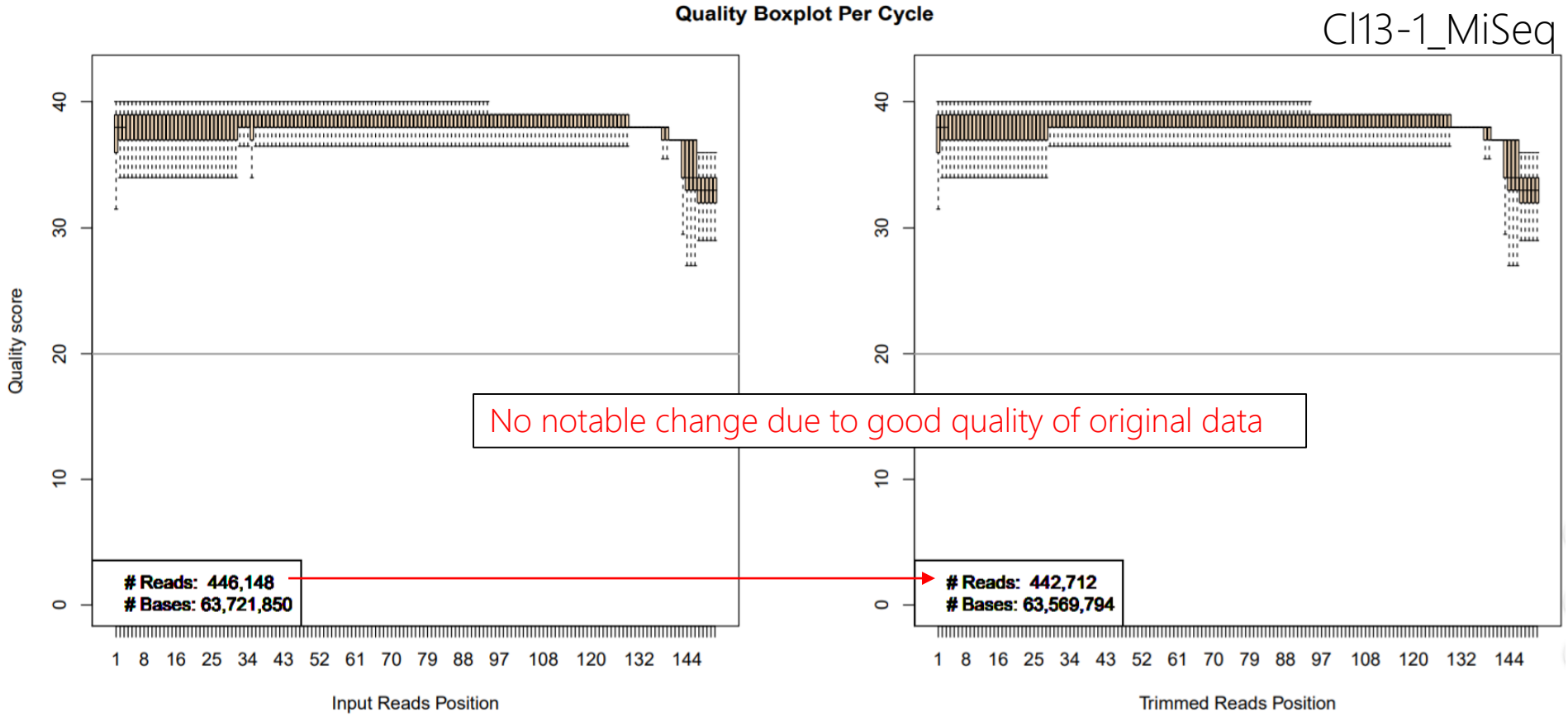
*Kmer : align unit is K bps, not read length

- Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data (2015)

3. Quality Filter

Objective | Removing sequencing reads with imperfect quality

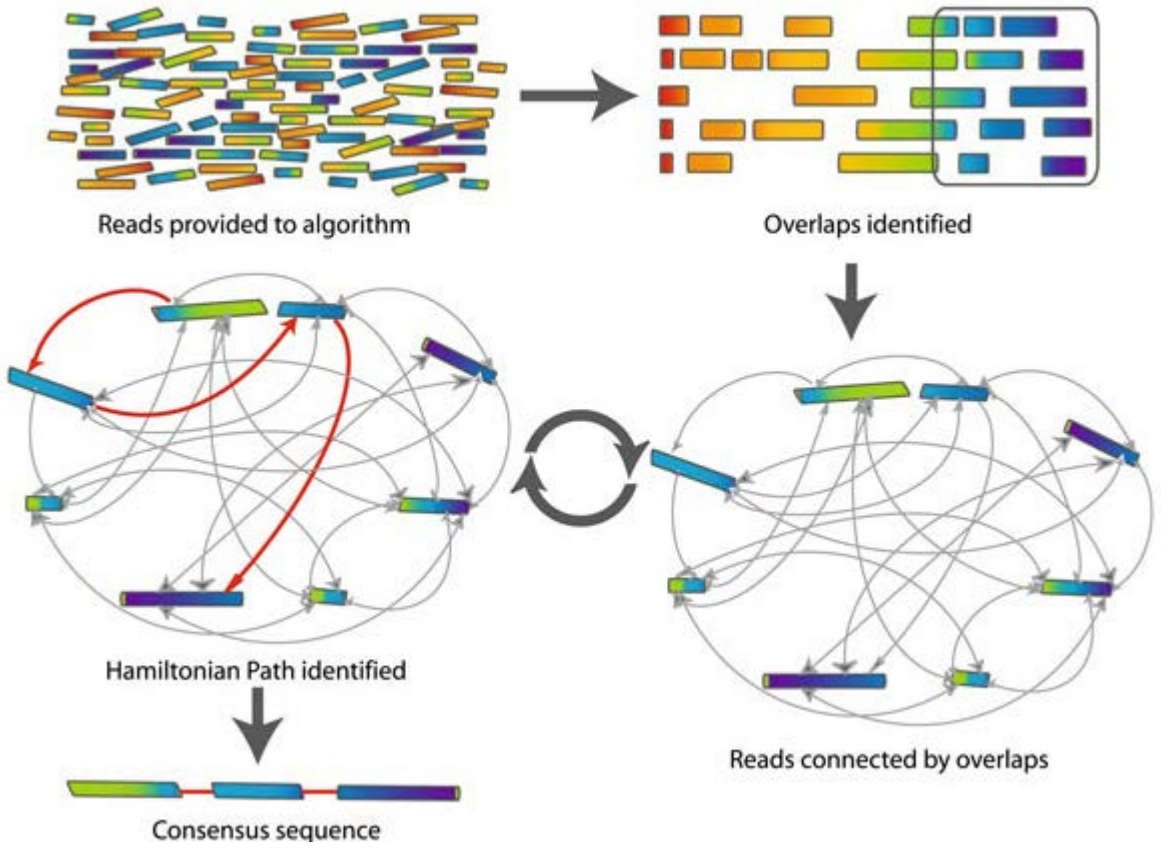
Program | FaQCs (2014)



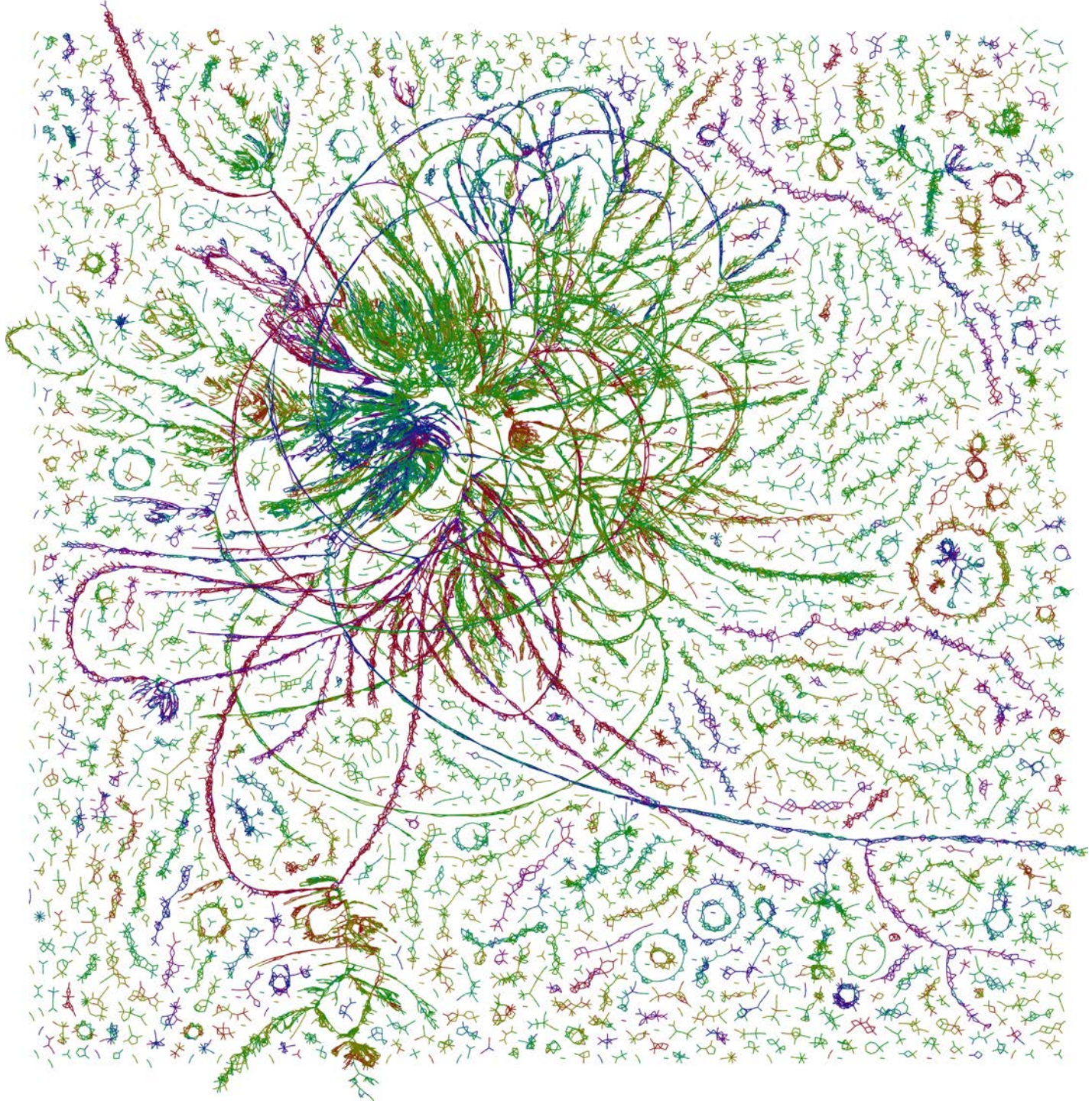
4. *De novo* Assembly

Objective | Concatenating refined reads into contigs

Program | SPAdes (2012)



(contig)



Difference in capabilities of *de novo* assembler

5. Screening

Objective | Screening assembled contigs against reference sequences database

Program | BLAST+ (1990) – megablast, discontiguous megablast, blastn, tblastx

Reference DB | NCBI RefSeq Viral sequences

algorithms	sequence type	similarity	sensitivity	note
megablast	DNA	High	Low	more than 95% identity
discontiguous megablast	DNA	Somewhat	High	ignoring 3rd base mismatches
blastn	DNA	Somewhat	Somewhat	allow least similarity
tblastx	Amino acid			Bottleneck stage

Reference Database NCBI RefSeq Viral sequences + Lab’ s own reference sequences

NCBI RefSeq : non-redundant set of reference sequences database provided by NCBI

- Updates in odd number months (January, March, May, July, September, November)
- 7546 species included (Latest version)

Lab’ s own reference sequences : reference sequences which are not included in RefSeq

- 29 Sequences
- Seeweis virus, Imjin virus, Muju virus, Seoul virus, Soochong virus, etc.

Final product BLAST result comes out as file (viewed and modified in Microsoft Excel)

① Contig No	② Contig Len.	③ Subject Accession	E-Value	Bitscore	④ Align Len.	⑤ Subject Title	⑥ Status	⑦ Identity	⑧ Start	⑨ End
NODE_9	2458	NC_025352	0	1618	2272	Mojiang virus isolate Tongguan1	complete	76.012	13750	11482
NODE_1	2185	NC_025352	0	1741	2113	Mojiang virus isolate Tongguan1	complete	78.325	16965	14855
NODE_28	1638	NC_025352	0	1229	1564	Mojiang virus isolate Tongguan1	complete	77.621	10617	9059

① Contig Number : ID of screened contig among assembled ones

② Contig Length : Length of screened contig

③ Accession Number : NCBI accession number of reference sequence

④ Alignment Length : Length of aligned sequences between contig and reference sequence

⑤ Subject Title : Taxonomy, Strain title of reference sequence

⑥ Status : Normally complete, seldom partial

⑦ Identity : Identical sequences / Alignment length * 100

⑧ Start : Starting sequence of aligned sequences in reference sequence

⑨ End : Ending sequence of aligned sequences in reference sequence

Raw score(S) : degree of homology in a certain alignment

"Unless the scoring system is understood, citing a raw score alone is like citing a distance without specifying feet, meters, or light years"

- BLAST manual

Bitscore : normalized raw score to compare between different scoring systems

$$\text{Bitscore} = \frac{\lambda S - \ln K}{\ln 2}$$

λ, K : statistical factor

Higher is better, but don't adequately represent significance of alignment

E-value : number of expected hits of similar quality (score) that could be found just by chance

$$\text{E-value} = \frac{m \; X \; n}{2^{Bitscore}}$$

m : query sequence length

n : database sequences length

Lower is better. Especially, smaller than 1 x e^{-50} means very high quality match (empirically)

Published papers use E-value 1 x e^{-6} or 1 x e^{-5} as a criterion

Characterizing the virome of Ixodes Ricinus ticks from northern Europe (2017), Scientific Reports

"screened with blastX against a local database with hits with an *e*-value of **1×10^{-5}** or better collated"

Virome analysis for identification of novel mammalian viruses in bats from Southeast China (2017), Scientific Reports

"In all the blast results, optimal results were used as the gene annotation with the parameter of E value < **$10e^{-5}$**"

Metagenomic Profile of the Viral Communities in Rhipicephalus spp. Ticks from Yunnan, China(2015), PLOS ONE

"Based on the most significant BLASTx similarities (e-value < **10^{-6}**)"

Comparison with EDGE

Strength

1) Easy to install and modify : few efforts to need when changing any program in pipeline
2) Easy to update or add database sequences
3) Can extract virus-like contigs
4) Can choose host whose genome will be subtracted from raw data

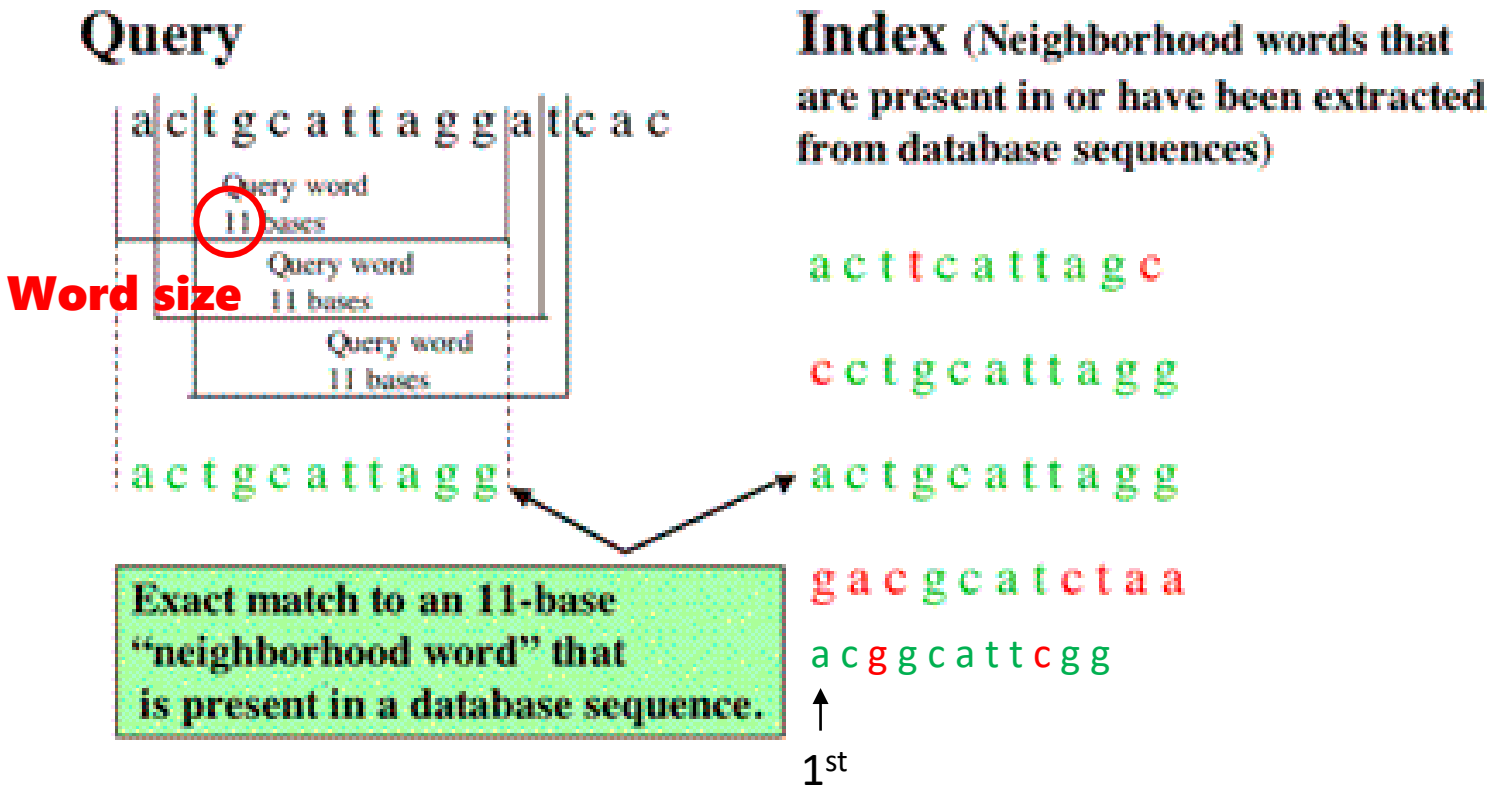
Weakness

1) No support for taxonomic classification
2) No support for GUI (Graphic User Interface)
3) Low-level pre-assembly filtering (repeat, low complexity, etc) ⇒ Perhaps, reason for too many results

Further advancement

1) Taxonomic classification
2) More pre-assembly filtering (repeat, low complexity, etc)
3) Development for GUI (Graphic User Interface)
4) Addition of Bacterial screening step

When mapping query to reference sequence, a certain size of sequences has to exactly match with each other, and the size is **Word size**



algorithms	similarity	Word size	sensitivity	note
megablast	High	28 bp	Low	more than 95% identity
discontiguous megablast	Somewhat	11 – 12 bp	High	ignoring 3rd base mismatches
blastn	Somewhat	7 – 11 bp	Somewhat	allow least similarity

But in case of discontinuous megablast, 3rd base mismatches are ignored when matching so as to consider wobble pairing

[Supplementary Slide 2] E-value

①

②

$$\text{E-value} = \frac{m\ X\ n}{2^{Bitscore}}$$

m : query sequence length

n : database sequences length

$\Rightarrow \text{E-value}_1 = 2 * \text{E-value}_2$

[Supplementary Slide 3] Duration

Environment | Workstation (CJ 식품안전관), Server (문숙의학관)

Sample contigs | Cl13-1 NGS data (total 98.7 MB for paired files), 7027

Workstation (CJ 식품안전관)

Step	Duration (seconds)
Adaptor Clipping	6
Host Removal	126
Quality Filter	162
De novo Assembly	237
blastn	113
megablast	2
dcmegablast	85
blastx	8929 (2h 28m)

Server (문숙의학관)

Step	Duration (seconds)
Adaptor Clipping	6
Host Removal	183
Quality Filter	288
De novo Assembly	256
blastn	128
megablast	2
dcmegablast	85
blastx	9946 (2h 46m)