Pilot study to evaluate the effect of different sequencing platforms and virus species on genome assembly quality

Amit Kumar Gupta and Manoj Kumar*
Bioinformatics Centre, CSIR-Institute of Microbial Technology
Sector 39 A, Chandigarh- 160036, India
Email: amitg@imtech.res.in, manojk@imtech.res.in

Introduction

➢The diverse viral species such as Dengue, HIV, Influenza, WNV, Hepatitis etc. are the key causative agent to the numerous diseases. Next Generation Sequencing (NGS) has an important role in the viral diagnostics

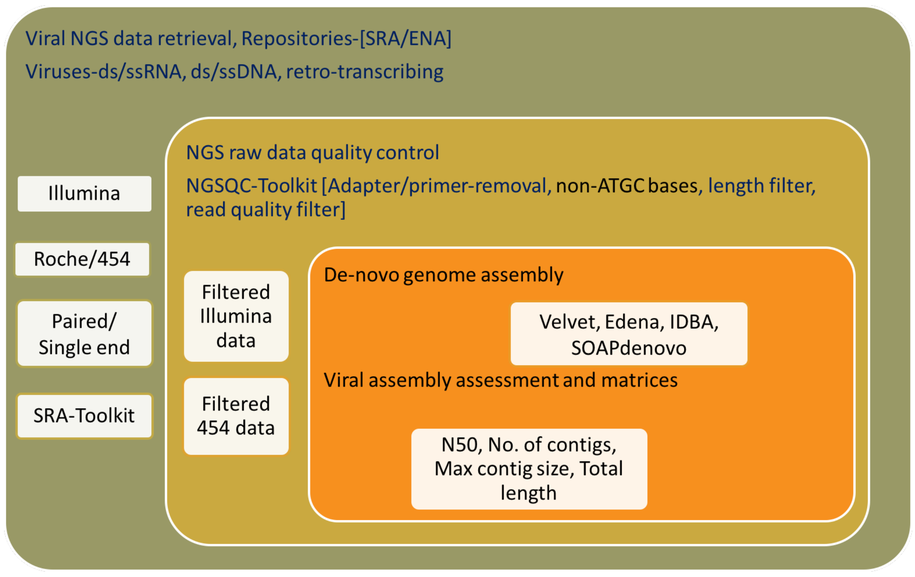
➢Along with the advancements of high throughput sequencing technologies, various de novo genome assembly tools have been evolved rapidly

➢There are various studies regarding the evaluation and comparison of genome assembly algorithms based on different criteria. However, these evaluations mainly focused on or utilized the NGS data from human, bacterial or plant origin.

➢There is scant attention has been given to the performance of assemblers on viral NGS data from different sequencing platforms

➢Briefly, in this study, we have analyzed the genome assembly from different assembly algorithms on the real NGS dataset of diverse viruses from different sequencing platforms i.e. of Illumina (GA II, GA IIx, HiSeq 2000, MiSeq) and 454/Roche (GS Junior, GS FLX, GS FLX Titanium)

Materials and Method



Results

Table 1. The illumina viral NGS data and quality analysis statistics

Viruses	Run Accession	Technology	Layout	Number of raw reads	Number of quality filtered reads	Reads length (*bp*)
Influenza virus A	*ERR045841*	*GA II*	*Paired*	*5,18,134*	*3,66,219*	*54*
Human herpesvirus 8	*ERR244026*	*GA IIx*	*Paired*	*2,955,212*	*1,784,389*	*76*
HIV 1	*SRR527726*	*HiSeq 2000*	*Paired*	*3,488,150*	*2,435,004*	*101*
Rhinovirus A	*SRR499802*	*HiSeq 2000*	*Paired*	*16,947*	*7,493*	*101*
Hepatitis C virus	*ERR118962*	*MiSeq*	*Paired*	*6,44,828*	*3,02,839*	*131*
Dengue virus 3	*SRR546416*	*MiSeq*	*Paired*	*4,72,546*	*15,800*	*225*
WNV	*SRR546546*	*MiSeq*	*Paired*	*1,61,067*	*881*	*225*
Hepatitis B virus	*DRR001353*	*GA IIx*	*Single*	*7,68,941*	*4,50,365*	*64*

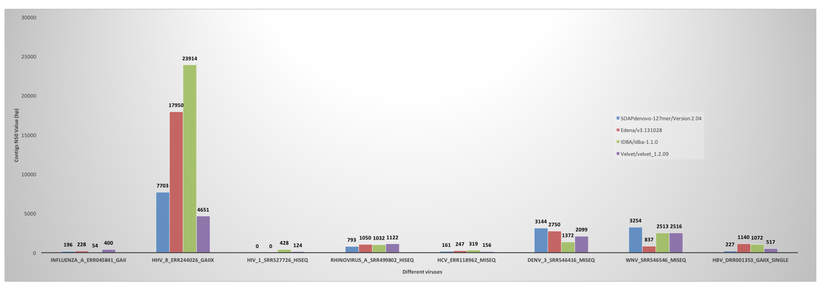


Figure 1. N50 value distribution of illumina data genome assemblies for different assemblers

Table 2. The Roche/454 Viral NGS Data Statistics and quality analysis

Viruses	Run Accession	Technology	Layout	Number of row reads	Number of quality reads
Influenza A virus (A/Rico)	*ERR247196*	*GS Junior*	*Single*	*8,277*	*4984*
Influenza virus A	*ERR180941*	*GS FLX Titanium*	*Single*	*19,577*	*17,750*
HIV 1	*SRR002785*	*GS FLX*	*Single*	*2,243*	*2127*
HIV 1	*ERR102556*	*GS FLX Titanium*	*Single*	*16,862*	*15175*
SIV	*SRR585831*	*GS Junior*	*Single*	*6,403*	*6,365*
Enterovirus C	*SRR038588*	*GS FLX*	*Single*	*1,63,500*	*1,45,356*
Yellow fever virus	*SRR038591*	*GS FLX*	*Single*	*31,824*	*27,563*
Mumps virus	*SRR038592*	*GS FLX*	*Single*	*11,289*	*7887*
Rotavirus	*SRR038593*	*GS FLX*	*Single*	*15,249*	*10,310*
Human herpesvirus 3	*SRR038595*	*GS FLX*	*Single*	*40,929*	*30.044*
WNV	*SRR331093*	*GS FLX Titanium*	*Single*	*3,465*	*1,386*
Dengue virus 2	*SRR534172*	*GS FLX Titanium*	*Single*	*2,339*	*1605*
Hepatitis C virus	*ERR149035*	*GS FLX Titanium*	*Single*	*71,134*	*55,268*

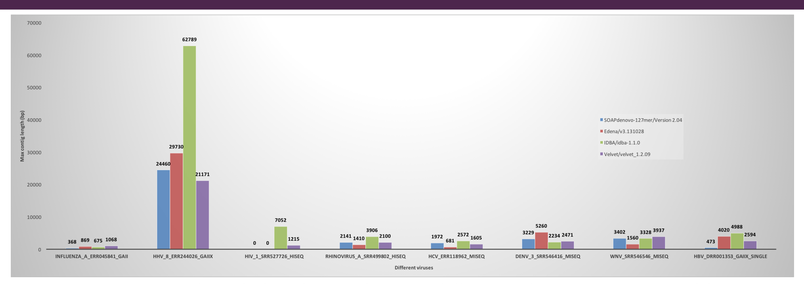


Figure 2. Max contig length distribution from viral illumina data genome assemblies

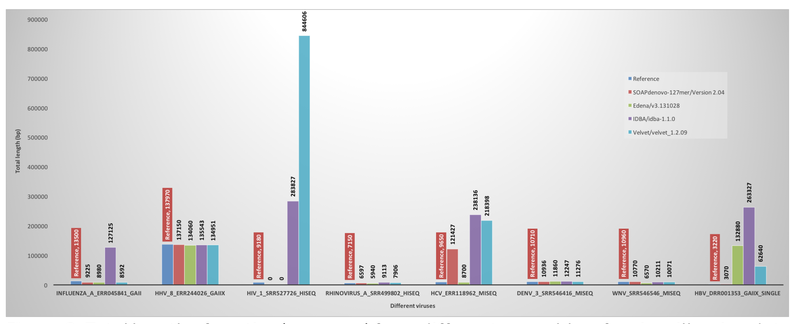


Figure 3. Total length of contigs (coverage) from different assemblers for virus illumina data

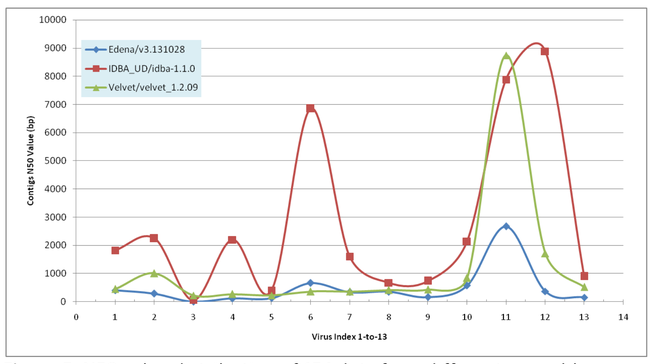


Figure 5. N50 value distribution of 454 data from different assemblers

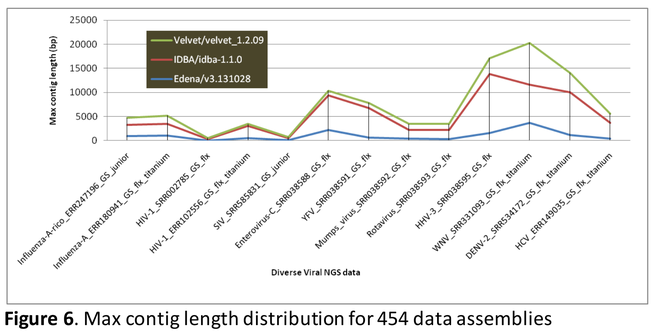
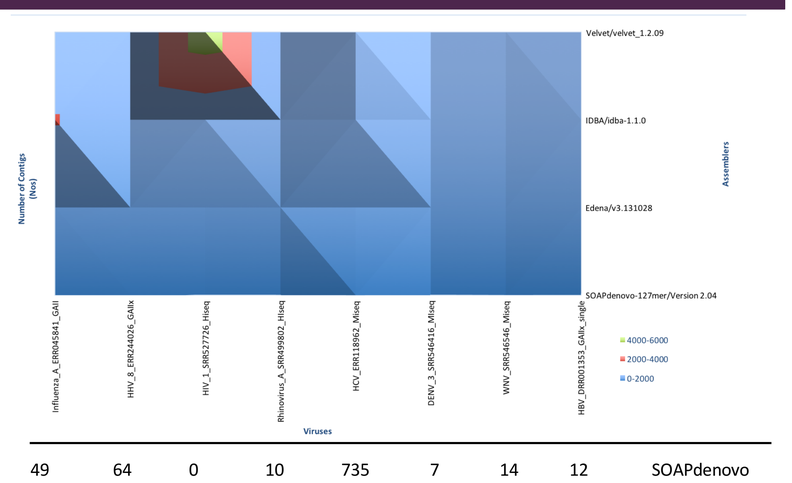


Figure 6. Max contig length distribution for 454 data assemblies



49	64	0	10	735	7	14	12	SOAPdenovo
48	16	0	8	32	7	10	293	Edena
2134	42	650	30	1014	17	6	319	IDBA
28	51	5180	14	1373	10	5	146	Velvet

Figure 4. Contour graph showing number of contigs from Illumina data assemblies

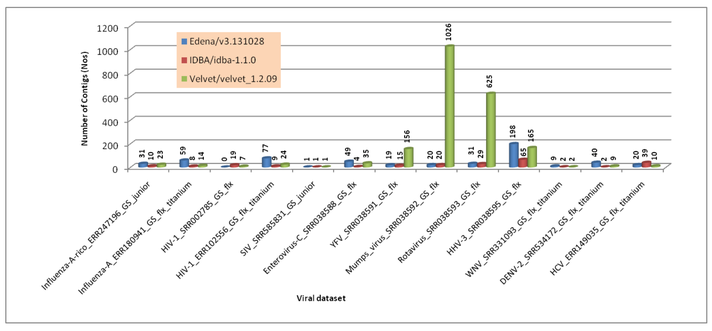


Figure 7. Bar graph showing number of contigs from viral 454 data assemblies

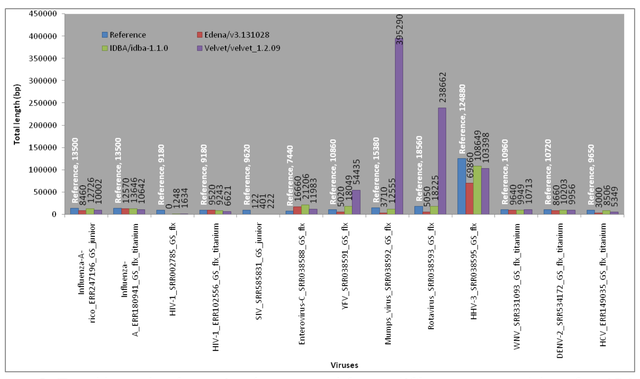


Figure 8. Total length of contigs (coverage) from different assemblers for 454 data

Conclusions

❖We have analyzed the genome assembly and performance of assembler based on different criteria i.e. contig numbers, N50 statistics, maximum contig length, genome coverage etc.

❖We have Analyzed the different viral categories i.e. ssRNA (+ve) viruses (like Hepatitis C virus, Dengue, WNV etc.), ssRNA (-ve) viruses (Eg. Influenza viruses), Retro-transcribing viruses (like HIV, SIV, Hepatitis B virus etc.), dsRNA viruses (Eg. Rotavirus), dsDNA viruses (Eg. Human herpes virus) using widely employed assemblers i.e. Velvet, SOAPdenovo, IDBA, Edena etc.

❖We have observed that existing assemblers are inconsistent and perform poor on retro-viral NGS data

❖Further, there is need to analyze misassemblies, and to extend the study to evaluate some viral specific assemblers along with the other most common assemblers

References

1. Alkan C, *et al.* (2011). Limitations of next-generation genome sequence assembly. *Nature methods*, 8(1):61-65.
2. Miller JR, *et al.* (2010). Assembly algorithms for next-generation sequencing data. *Genomics,* 95: 315–327.
3. Earl DA, *et al.* (2011). Assemblathon 1: A competitive assessment of *de novo* short read assembly methods. *Genome Res.,* 21:2224-2241.
4. Bradnam KR, *et al*. (2013). Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *Gigascience*., 22;2(1):10.
5. Bao S, *et al. (2011).* Evaluation of next-generation sequencing software in mapping and assembly. *J of Human Genetics,* 56: 406-414.
6. Zhang W, *et al.* (2011). A Practical Comparison of *De Novo* Genome Assembly software Tools for Next Generation Sequencing Technologies. *PLoS ONE,* 6(3): e17915.
7. Beerenwinkel N, *et al*. (2012). Challenges and opportunities in estimating viral genetic diversity from next-generation sequencing data. *Front Microbiol*., 3: 329.
8. Salzberg SL, *et al.* (2012). GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res.,* 22:557-567.
9. Magoc T, *et al*. (2013). GAGE-B: an evaluation of genome assemblers for bacterial organisms. *Bioinformatics,* 29(14):1718-25.
10. Barzon L, *et al*. (2012). Applications of next-generation sequencing technologies to diagnostic virology. *J Gen Virol.*, 93(9): 1853–1868.
11. Zerbino DR and Birney E (2008). Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.*, 18: 821-829.
12. Zerbino DR, (2010). Using the Velvet *de novo* assembler for short-read sequencing technologies. *Curr Protoc Bioinformatics*.
13. Peng, Y., *et al*. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth, *Bioinformatics*, 28, 1420-1428.
14. Hernandez D, *et al*. (2008). De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. *Genome Res.*18(5):802–809.

Acknowledgements

Council of Scientific and Industrial Research (CSIR), India and Department of Biotechnology (DBT), Government of India.

IMTechCon 4th - 6th October 2017 An Industry-Academia Meet