NGS for Studying Viruses "Beyond the Consensus"

Jan T. Kim
The Pirbright Institute, Pirbright, United Kingdom
Kim JT (2014) *EMBnet.journal* **20**(Suppl A), e774. http://dx.doi.org/10.14806/ej.20.A.774

High mutation rates in viruses (especially RNA viruses) have profound consequences on viral evolution, including the formation of quasispecies. These have long been studied in theory (Eigen, 1971) and in silico (Wilke *et al.*, 2001). NGS technologies provide new opportunities to directly observe sequence diversity and its evolution in viruses (Wright *et al.*, 2011) and other systems (Schütze *et al.*, 2011).

Profiles of base frequencies can be constructed from virus NGS data. They are used in a number of well established bioinformatics contexts, including DNA binding site representation (Stormo, 2000) and progressive multiple alignment (Larkin *et al.*, 2007). Profiles can be formalised as elements of a continuous sequence space (Vingron and Sibbald, 1993), and they serve as a basis for information theoretic analyses (Schneider *et al.*, 1986; Kim *et al.*, 2003) and statistical learning approaches (Kim *et al.*, 2004).

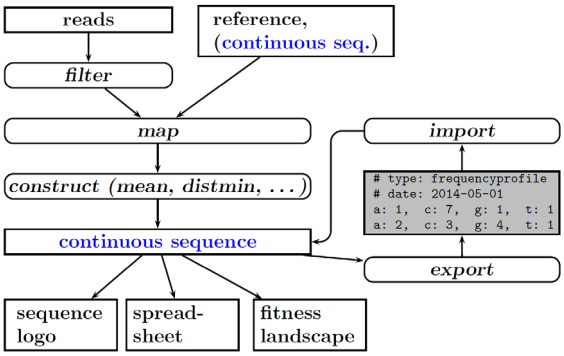

Figure 1. Continuous sequences as a point of departure for many bioinformatic analyses.

Given this basis, profiles provide a point of departure for many types of analyses of NGS data comprising diverse populations, illustrated in Figure 1. I will demonstrate their construction, outline some opportunities for their future use in studying viral diversity and quasispecies, and discuss technical requirements for their appropriate and efficient use.

References

Eigen M (1971) Selforganization of matter and the evolution of biological macromolecules. *Naturwissenschaften* **58**, 465–523. http://dx.doi.org/10.1007/BF00623322

Kim JT, Martinetz T, and Polani D (2003) Bioinformatic principles underlying the information content of transcription factor binding sites. *J Theor Biol* **220**, 529–544. http://dx.doi.org/10.1006/jtbi.2003.3153

Kim JT, Gewehr JE, and Martinetz T (2004) Binding matrix: A novel approach for binding site recognition. *J Bioinform Comput Biol* **2**, 289–307. http://dx.doi.org/10.1142/S0219720004000569

Larkin M, Blackshields G, Brown N, Chenna R, McGettigan P, et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947–2948. http://dx.doi.org/10.1093/bioinformatics/btm404

Schneider TD, Stormo GD, Gold L, and Ehrenfeucht A (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol* **188**, 415–431. http://dx.doi.org/10.1016/0022-2836(86)90165-8

Schütze T, Wilhelm B, Greiner N, Braun H, Peter F, et al. (2011) Probing the SELEX process with next-generation sequencing. *PLoS One* **6**, e29604. http://dx.doi.org/10.1371/journal.pone.0029604

Stormo GD (2000) DNA binding sites: Representation and discovery. *Bioinformatics* **16**, 16–23. http://dx.doi.org/10.1093/bioinformatics/16.1.16

Vingron M and Sibbald PR (1993) Weighting in sequence space: A comparison of methods in terms of generalized sequences. *Proc Natl Acad Sci U S A* **90**, 8777–8781. http://dx.doi.org/10.1073/pnas.90.19.8777

Wilke CO, Wang JL, Ofria C, Lenski RE, and Adami C (2001) Evolution of digital organisms at high mutation rate leads to survival of the flattest. *Nature* **412**, 331–333. http://dx.doi.org/10.1073/pnas.90.19.8777

Wright CF, Morelli MJ, Thébaud G, Knowles NJ, Merzyk P, et al. (2011) Beyond the consensus: Dissecting within-host viral population diversity of foot-and-mouth disease virus by using next-generation genome sequencing. *J Virol* **85**, 2266–2275. http://dx.doi.org/10.1128/JVI.01396-10