Comparative genomics of nidoviruses: towards understanding the biology and evolution of the largest RNA viruses

Gulyaeva, A.

Summary

Sumenvatting

List of abbreviations

Curriculum Vitae

List of publications

Acknowledgements

SUMMARY

The order *Nidovirales* is a monophyletic group of positive-sense single-stranded RNA viruses that infect vertebrate and invertebrate hosts, and include viruses with largest RNA genomes. A set of hallmark characteristics distinguish nidoviruses from other RNA viruses: genome organization, mechanisms of genome expression, a synteny of conserved replicative domains. Only a few selected nidoviruses are subject of comprehensive experimental research. At the same time, the advent of next generation sequencing has greatly accelerated the rate of nidovirus discovery. As a result, genome sequence is the only characteristic available for a large and ever growing share of nidoviruses. These developments determine the key role of comparative genomics in further nidovirus characterization. Comparative genomics identifies homologous regions of genomes and proteins, facilitating evolutionary studies, and functional and structural characterization of newly discovered and already known viruses. Specifically, it promotes transfer of functional annotation from experimentally characterized viruses and hosts to newly discovered virus genomes, and defines constraints of natural variation for all viruses, including experimentally characterized. In this thesis, we used comparative genomics to characterize various aspects of nidovirus biology and evolution. This study was conducted in collaboration with other researchers, who discovered new viruses and sequenced their genomes (Chapters 2 and 4), or characterized virus proteins experimentally following bioinformatics sequence analysis (Chapter 3). **Chapter 1** provides background on nidoviruses and techniques of comparative genomics available by the end of 2014, when the project that resulted in this thesis started. **Chapter 2** describes characterization of arterivirus polyprotein 1ab N-terminus encoding multiple papain-like proteases. The analysis relied on previous research on this region and included 5'-terminus of the divergent wobbly possum disease virus genome, sequencing of which was completed as part of the study. The study offers insight into the role and contribution of gene duplication to nidovirus adaptation. **Chapter 3** presents discovery of the fifth replicative domain universally conserved in all nidoviruses, nidovirus RdRp-associated nucleotidyltransferase or NiRAN. NiRAN conservation in nidoviruses, its evolutionary origin, biochemical activity and potential function were analyzed. **Chapter 4** focuses on discovery and characterization of a highly divergent nidovirus with the largest known RNA genome, planarian secretory cell nidovirus or PSCNV. Both unique and conserved features of its genome, proteome and expression were revealed in this study. Moreover, PSCNV discovery advanced our understanding of RNA genome expansion limits. **Chapter 5** addresses an important technical challenge of nidovirus comparative genomics. Proteomes of RNA viruses, including nidoviruses, are dominated by large multidomain polyproteins, although standard tools for homology detection were trained on single-domain proteins. Consequently, homologous relationships of domains in polyproteins may

remain undetected due to underestimation of hits statistical significance. To mitigate this problem, we introduced a tool, called LArge Multidomain Protein Annotator or LAMPA, that gradually splits polyprotein sequence into smaller queries in a biologically reasonable manner, improving estimation of hits statistical significance and annotation coverage. **Chapter 6** discusses how discoveries of recent years, including the ones described in this thesis, advanced our understanding of two fundamental aspects of nidovirus biology. First, we reexamine nidovirus hallmarks, prompted by discovery of novel and divergent nidoviruses and their bioinformatics analysis. Second, we review new insights into the mechanisms of large scale sequence change in nidovirus genomes, which have the largest RNA genome size range.