Rapid analysis of SARS-CoV-2 genomic content

Kristen L. Beck*, Simone Bianco, Gowri Nayar, Harsha Krishnareddy, Akshay Agarwal, Hakan Bulu, James Kaufman, Vandana Mukherjee, Edward Seabolt

AI and Cognitive Software, IBM Almaden Research Center
*Correspondence: klbeck@us.ibm.com

During this global pandemic, SARS-CoV-2 sequencing efforts have ramped up to address the needs of this once in lifetime public health emergency. Hundreds of new sequences have been deposited from around the world. We have retrieved all global SARS-CoV-2 genome sequences and present the genes, proteins, and functional domains contained therein. This *in silico* annotation yielded 231,382 sequences across these biological entities some of which are conserved and others represent emerging variants. From this collection of 1,885 SARS-CoV-2 genomes (retrieved 2020-03-28), we identify a core set of 11 genes. Each of these is present with 217 average variants even in this small collection of viral genomes (number of variants range 3–811). For proteins, the magnitude is comparable with an average of 172 variants per protein name (range 3–659). Furthermore, we identify enriched motifs present in the spike glycoprotein, a significant gatekeeper in host cell invasion. As part of IBM's COVID-9 response, we release the derived gene, protein, and domain data to the scientific and medical communities.