Title: Respiratory Syncytial Virus whole-genome sequencing identifies convergent evolution of sequence duplication in the C-terminus of the G gene.

Seth A. Schobel¹,²,⁷, Karla M. Stucker¹, Martin L. Moore³, Larry J. Anderson³, Emma K. Larkin⁵,⁶, Jyoti Shankar¹, Jayati Bera¹, Vinita Puri¹, Meghan H. Shilts¹, Christian Rosas-Salazar⁴, Rebecca A. Halpin¹, Nadia Fedorova¹, Susmita Shrivastava², Timothy B. Stockwell², R. Stokes Peebles⁵,⁶, Tina V. Hartert⁵,⁶, Suman R. Das¹*

¹Infectious Diseases and ²Bioinformatics Group, J. Craig Venter Institute, Rockville, MD
³Division of Infectious Diseases, Department of Pediatrics, Emory University School of Medicine and Children’s Healthcare of Atlanta, Atlanta, GA
⁴Division of Allergy, Immunology, and Pulmonary Medicine, Department of Pediatrics, Vanderbilt University School of Medicine, Nashville, TN
⁵Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN
⁶Division of Allergy, Pulmonary, and Critical Care Medicine, Department of Medicine, Vanderbilt University School of Medicine, Nashville, TN
⁷Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD

Key Words: Respiratory Syncytial Virus, next-generation sequencing, evolution, disease severity

Journal: Scientific Reports

Corresponding author:
Suman Ranjan Das
Infectious Disease Group
J. Craig Venter Institute
Rockville, Maryland 20850
Phone: 301-795-7328
Fax: 301-795-7070
E-mail: sdas@jcvi.org
SUPPLEMENTAL FIGURE LEGEND

Figure S3. Bayesian maximum clade credibility (MCC) trees for all available full G gene sequences downloaded from GenBank and down-sampled to include representative centroid sequences from 98% sequence identity gene clusters. The RSV-A G gene phylogeny (A) shows that relatively little new diversity is added compared with the whole-genome analyses. However, the RSV-B G gene phylogeny (B) shows that additional diversity is being sampled by including additional G gene sequences compared with the whole-genome analyses. This indicates better surveillance of RSV-B strains from G gene sequencing than from the available whole-genome dataset. Strain names are colored by the presence (red) or absence (blue) of the large G gene duplication, with study samples in darker shades of red and blue. Bayesian posterior probability > 0.9 are provided for key nodes.