Serotype diversity and putative novel *cps* loci in a large, global collection of pneumococci.

GPS
Global Pneumococcal Sequencing Project

Andries J. van Tonder[1], Rebecca A. Gladstone[1], Stephanie W. Lo[1], Moon H. Nahm[2], Anne von Gottberg[3], Martin Antonio[4], Dean B. Everett[5], Keith P. Klugman[6], Robert F. Breiman[7], Lesley McGee[8], Stephen D. Bentley[1] and the Global Pneumococcal Sequencing Consortium

[1]Infection Genomics, The Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SA, UK, [2]Division of Pulmonary, Allergy and Critical Care Medicine, Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA, [3]Centre for Respiratory Diseases and Meningitis, National Institute for Communicable Diseases, Johannesburg, South Africa, [4]Vaccines and Immunity Theme, The Medical Research Council Unit, Serekunda, The Gambia, [5]Malawi Liverpool Wellcome Trust Clinical Research Programme, Blantyre, Malawi [6]Hubert Department of Global Health, Rollins School of Public Health, Emory University, Atlanta, GA 30322, USA, [7]Emory Global Health Institute, Emory University, Atlanta, USA, [8]Respiratory Diseases Branch, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA,

Background: The pneumococcus produces a polysaccharide capsule, encoded by the *cps* locus, that provides protection against phagocytosis and determines serotype; nearly 100 serotypes have been identified with new serotypes still being discovered. The Global Pneumococcal Sequencing (GPS) project was established to sequence ~20,000 pneumococci from around the world to examine the effect of global vaccine introduction. The aims of this study were to: investigate the *cps* loci of 18,384 GPS genomes, characterise the diversity of the 66 serotypes included, and identify putative novel *cps* loci for further study.

Dataset: 18,384 genomes from 54 countries and 14 supraregions were included in this study (Figure 1). Approximately 60% of the isolates were from invasive disease whilst the rest were from carriage or unknown sources. A total of 66 serotypes and *cps* variants were included in this study (Figure 2).

Methods: Serotypes were assigned using SeroBA and sequence reads were mapped back to *cps* locus references using BWA to create alignments which were used to construct phylogenies using FastTreeMP. Samples with divergent *cps* loci were investigated further.

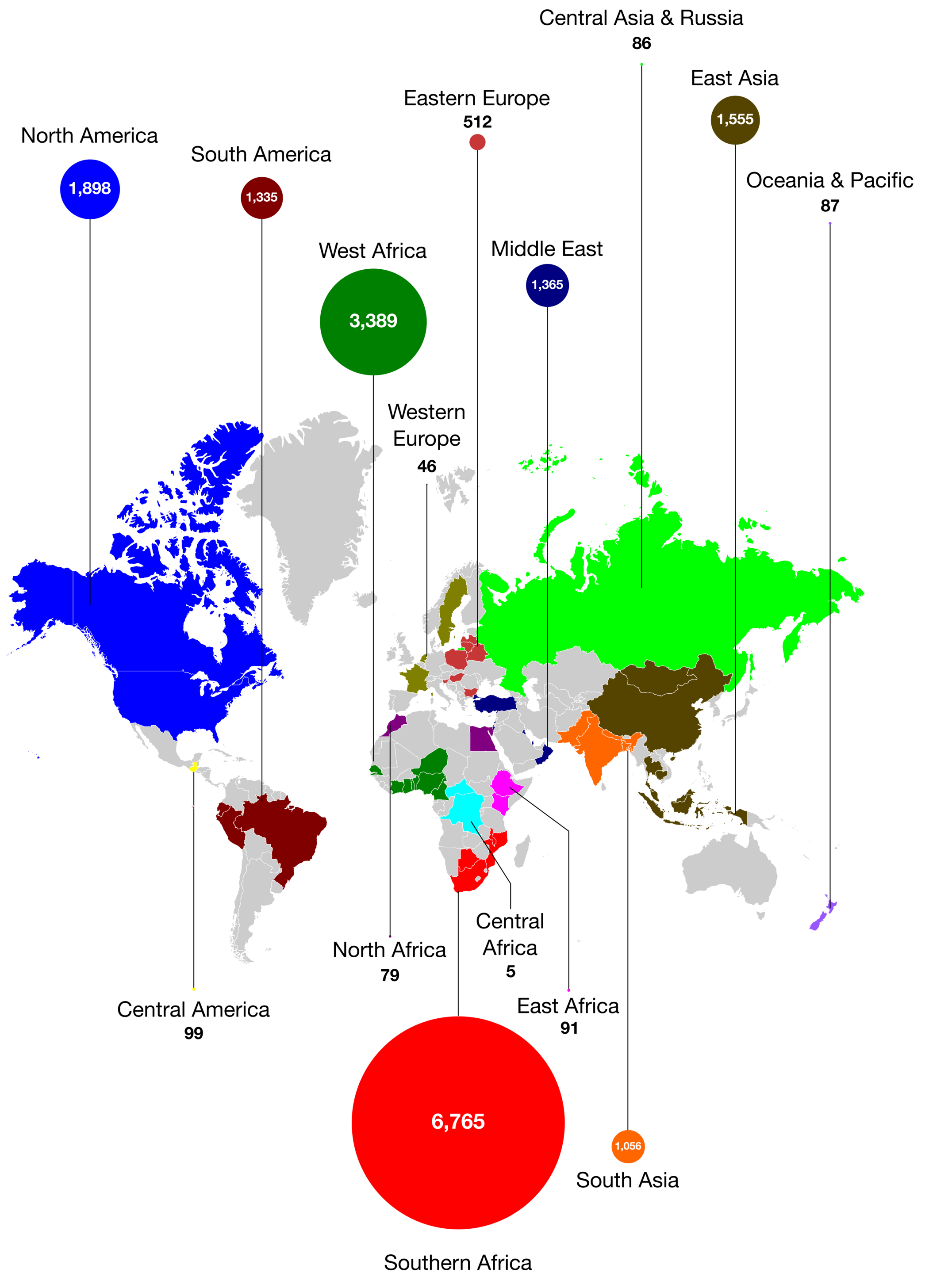


Figure 1: World map showing the site of isolation for each of the 18,384 genomes included in this study

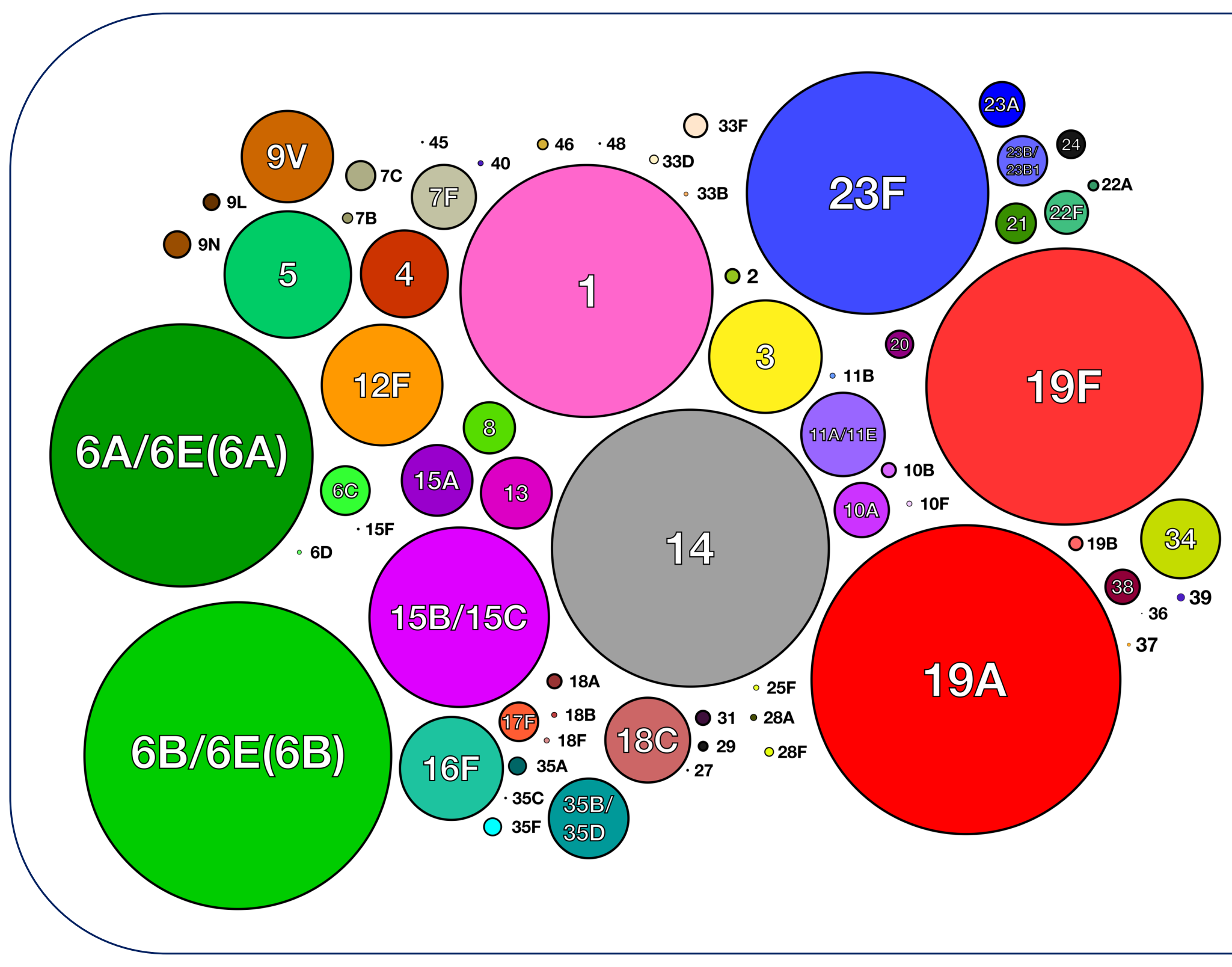
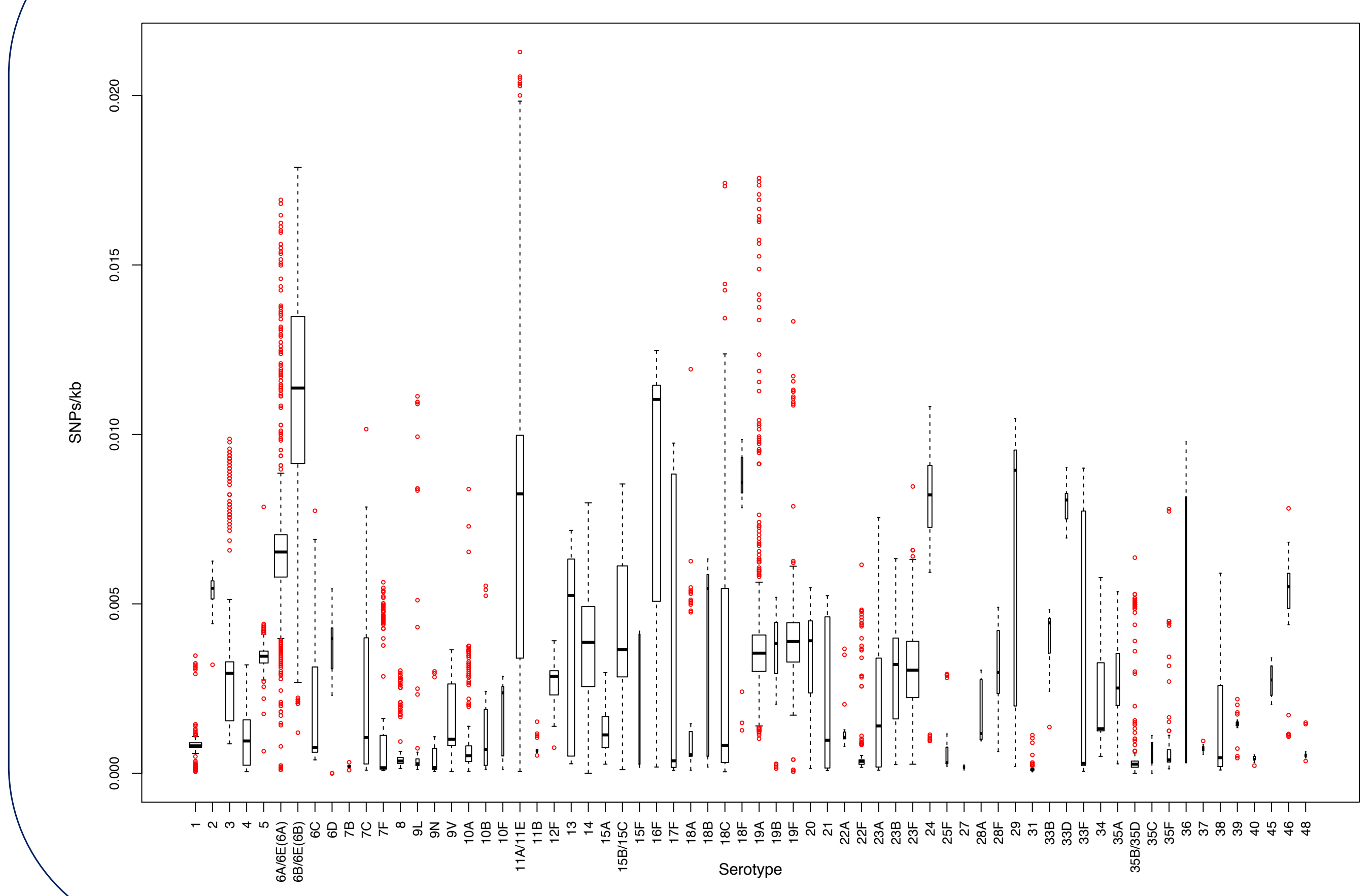


Figure 2: Bubble plot showing the 66 serotypes/*cps* variants included in this study. The area of each circle is proportional to the number of genomes included for that serotype (n = 18,384).



Results: Considerable diversity was observed in the 66 serotypes investigated (Figure 3) with some serotypes being highly conserved (serotype 1) and others very diverse (serotype 6A). Closer examination of each serotype identified eight putative novel *cps* loci: 9X, 11X, 16X, 18X, 18Xii, 29X, 33X and 36X which were found in 2.59% (476/18374) of the genomes. Three examples of the putative novel serotypes (9X, 18X and 29X) are depicted in Figures 4 to 6 (see below).

Figure 3: Boxplot showing *cps* locus diversity for 66 serotypes. The width of each boxplot is proportional to the square-root of the number of isolates. Outliers are highlighted in red.

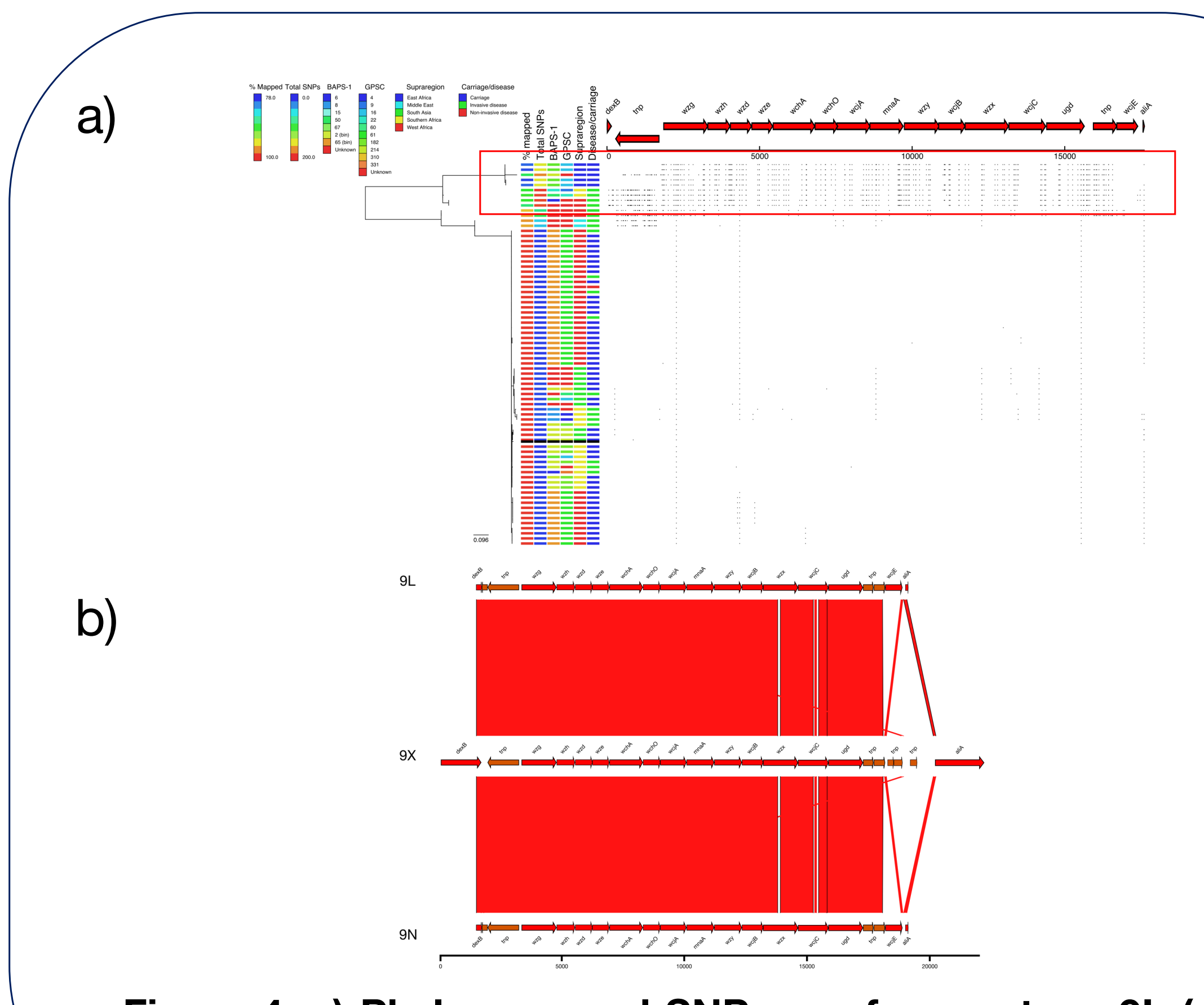


Figure 4: a) Phylogeny and SNP map for serotype 9L (n = 75). Putative serotype 9X *cps* loci are highlighted in red; b) ACT comparison of serotypes 9L, 9X and 9N.

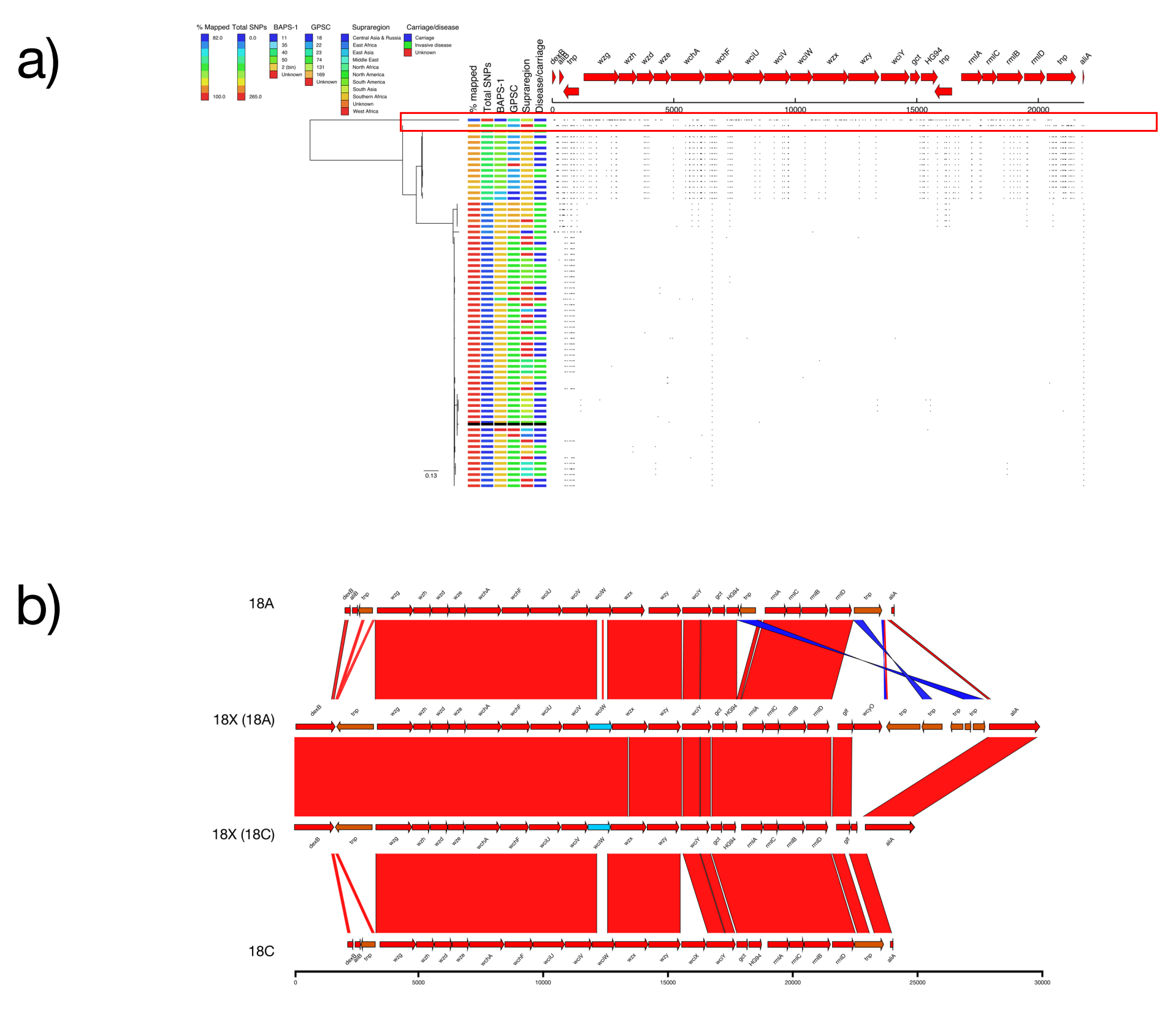


Figure 5: a) Phylogeny and SNP map for serotype 18A (n = 66). Putative serotype 18X *cps* loci are highlighted in red; b) ACT comparison of serotypes 18A, 18X and 18C.

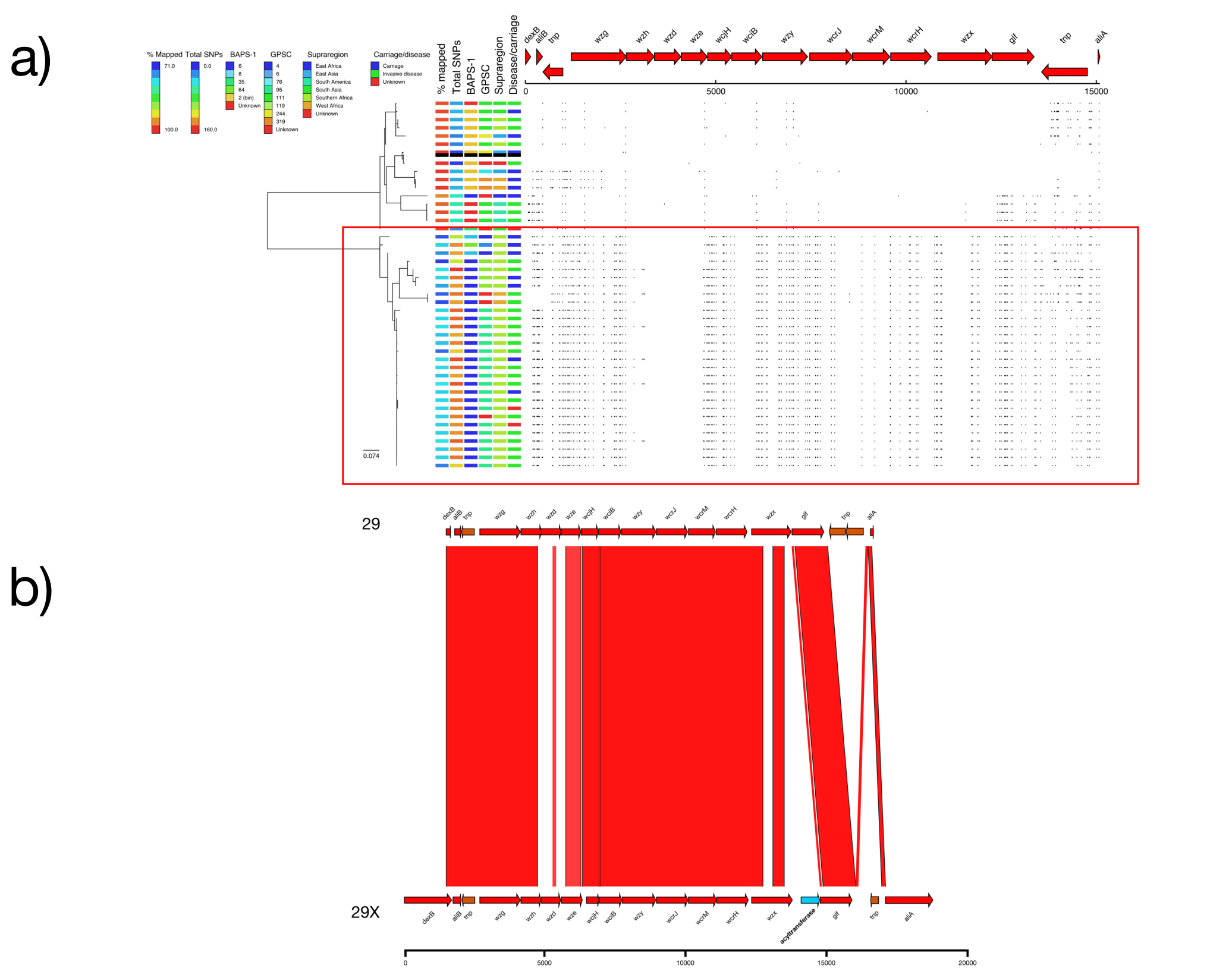


Figure 6: a) Phylogeny and SNP map for serotype 29 (n = 45). Putative serotype 29X *cps* loci are highlighted in red; b) ACT comparison of serotypes 29 and 29X.

Conclusions: The large number and global distribution of GPS genomes provided an unprecedented opportunity for identifying novel *cps* loci. Differing amounts of diversity were observed among the *cps* loci and eight putative novel *cps* loci were identified. Examples of each will be sent for structural and immunologic analysis.

BILL & MELINDA GATES foundation
wellcome sanger institute
CDC
EMORY
MRC The Gambia Unit
NATIONAL INSTITUTE FOR COMMUNICABLE DISEASES Division of the National Health Laboratory Service

Disclaimer: The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Centers for Disease Control and Prevention.