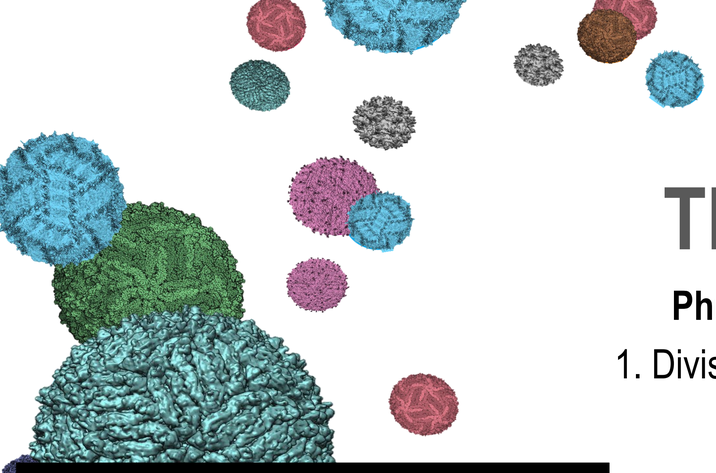


COMPARATIVE ANALYSIS OF THE CODON USAGE IN ZIKA VIRUS

Phuoc Truong [1] (phuoc.truong@utu.fi) and **Pere Puigbò** [1,2] (pedro.puigbo@utu.fi)

1. Division of Genetics and Physiology, Department of Biology, University of Turku, Finland

2. Turku Collegium for Science and Medicine, Turku, Finland

Turun yliopisto
University of Turku

ABSTRACT

Zika virus is a mosquito-borne Flavivirus that was first discovered in Uganda in 1947, from where it spread along the Equator through Asia. The first report of the virus in South America was in Easter Island (Chile) in 2014. Zika virus has been a major widespread pathogen in South America since early 2015. Currently, Zika viruses can be divided into three different genotypes: **East-African**, **West-African** and **Asian** (which also includes viruses circulating in North and South America).

In this study, we compare the synonymous codon usage among all three genotypes and within the Asian type. Full genome sequences (N=138) obtained from the NCBI's virus variation database are analyzed with the web-servers CAIcal and DendroUPGMA (D-UPGMA). The relative synonymous codon usage (RSCU) values are used to build a UPGMA dendrogram based on Pearson correlations. The resultant tree is free of any genetic recombination effect and clusters together those genomes with similar use of synonymous codons. The three major Zika genotypes can be easily identified from the tree, but there are only minor differences between sequences from Asia, Oceania and the Americas.

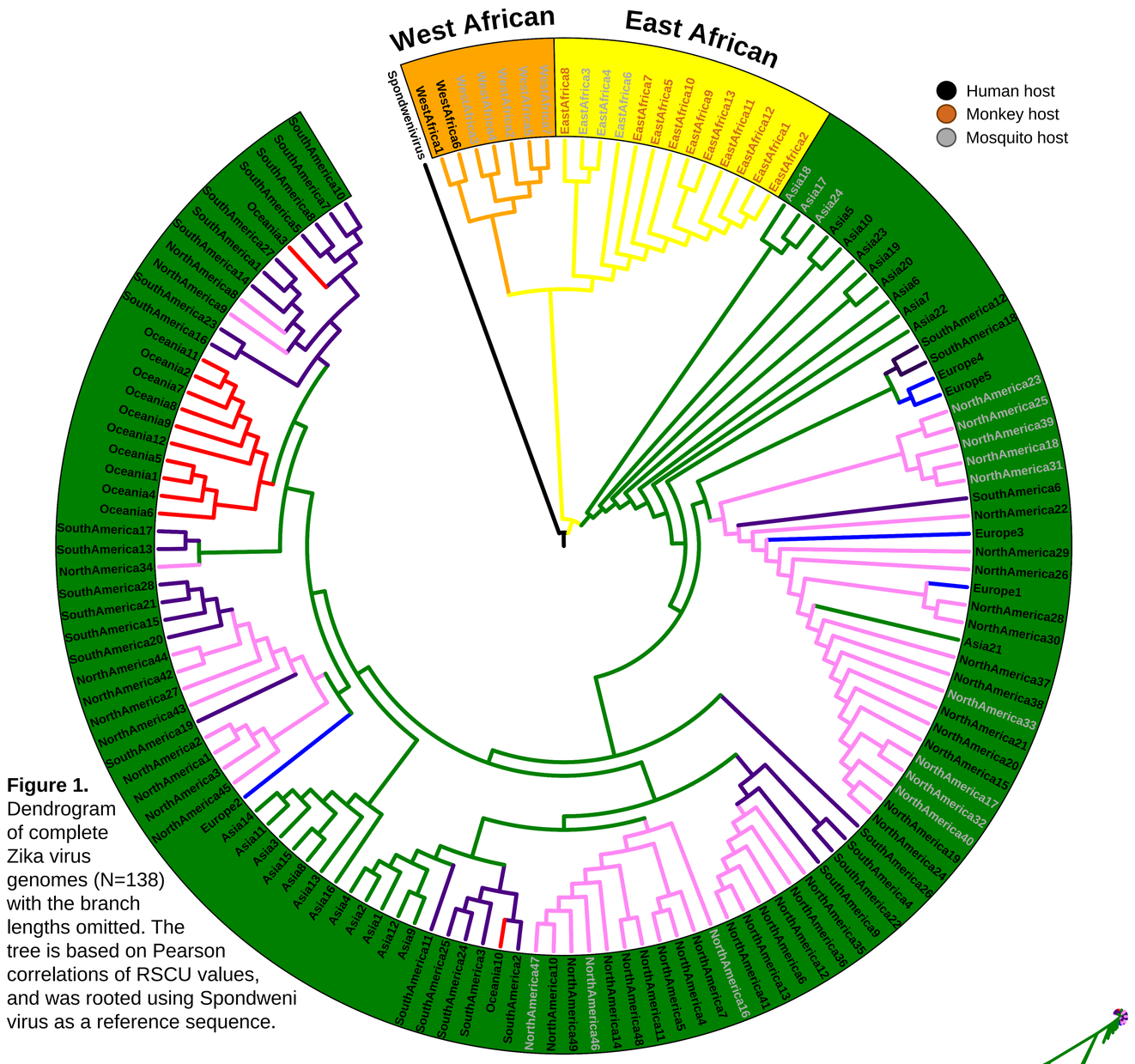


Figure 1. Dendrogram of complete Zika virus genomes (N=138) with the branch lengths omitted. The tree is based on Pearson correlations of RSCU values, and was rooted using Spondweni virus as a reference sequence.

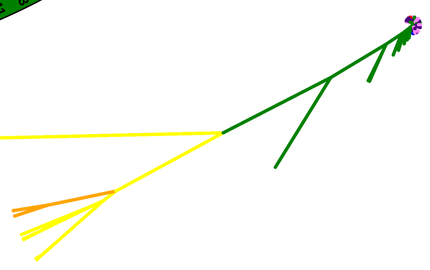


Figure 2. Dendrogram from figure 1. in unrooted format.

RESULTS

1. The results confirm the **Asian origin** of the recent outbreak in the Americas [1]. Cases detected in Europe were independently originated in **America**. Cases in Oceania can be traced to continental Asia. The most differentiation in codon usages between genotypes are highlighted in table 1.

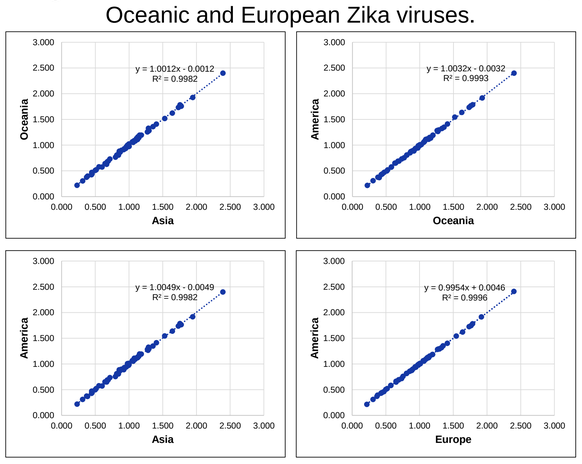
Table 1. Average and SD values of RSCUs in Zika virus.

		East Africa		West Africa		Asia		America	
Codon	AA	Average	StDev	Average	StDev	Average	StDev	Average	StDev
TTT	F	0.996	0.032	0.940	0.008	0.998	0.031	1.011	0.017
TTC	F	1.004	0.032	1.060	0.008	1.002	0.031	0.989	0.017
TTA	L	0.390	0.021	0.275	0.009	0.312	0.019	0.309	0.010
TTG	L	1.319	0.023	1.391	0.029	1.289	0.038	1.325	0.019
CTT	L	0.867	0.029	0.844	0.036	0.800	0.058	0.754	0.025
CTC	L	1.030	0.010	0.979	0.030	0.977	0.037	1.007	0.020
CTA	L	0.595	0.025	0.721	0.010	0.678	0.018	0.688	0.024
CTG	L	1.799	0.028	1.790	0.043	1.945	0.042	1.917	0.022
ATT	I	0.811	0.027	0.733	0.027	0.888	0.055	0.897	0.016
ATC	I	1.286	0.048	1.310	0.028	1.126	0.039	1.120	0.019
ATA	I	0.903	0.027	0.957	0.005	0.986	0.020	0.983	0.012
GTT	V	0.817	0.023	0.821	0.009	0.853	0.035	0.852	0.012
GTC	V	1.025	0.009	0.974	0.013	1.133	0.054	1.135	0.016
GTA	V	0.433	0.049	0.411	0.013	0.371	0.021	0.376	0.011
GTG	V	1.726	0.040	1.794	0.009	1.643	0.035	1.637	0.010
TCT	S	0.873	0.055	0.660	0.011	0.856	0.028	0.886	0.017
TCC	S	0.898	0.066	1.126	0.005	0.995	0.025	0.974	0.015
TCA	S	1.486	0.019	1.401	0.008	1.530	0.020	1.544	0.009
TCG	S	0.526	0.028	0.584	0.017	0.385	0.019	0.368	0.010
AGT	S	1.079	0.055	1.100	0.016	0.961	0.021	0.944	0.021
AGC	S	1.139	0.053	1.130	0.013	1.273	0.022	1.284	0.021
CCT	P	0.765	0.052	0.665	0.028	0.669	0.041	0.651	0.020
CCC	P	1.060	0.048	1.155	0.014	1.131	0.045	1.136	0.019
CCA	P	1.860	0.013	1.864	0.014	1.756	0.054	1.785	0.013
CCG	P	0.315	0.017	0.316	0.055	0.445	0.038	0.428	0.013
ACT	T	0.932	0.023	0.888	0.010	0.995	0.029	0.997	0.013
ACC	T	1.107	0.030	1.107	0.018	1.147	0.026	1.150	0.011
ACA	T	1.608	0.016	1.679	0.002	1.406	0.018	1.411	0.012
ACG	T	0.353	0.008	0.326	0.011	0.452	0.026	0.441	0.012
GCT	A	1.102	0.042	1.219	0.043	1.112	0.038	1.115	0.015
GCC	A	1.332	0.043	1.263	0.033	1.296	0.036	1.295	0.014
GCA	A	1.173	0.062	1.219	0.004	1.093	0.016	1.091	0.010
GCG	A	0.394	0.056	0.300	0.014	0.500	0.010	0.499	0.012
TAT	Y	0.845	0.071	0.802	0.024	0.717	0.038	0.734	0.018
TAC	Y	1.155	0.071	1.198	0.024	1.283	0.038	1.266	0.018
CAT	H	0.933	0.053	1.066	0.011	0.818	0.037	0.809	0.015
CAC	H	1.067	0.053	0.934	0.011	1.182	0.037	1.191	0.015
CAA	Q	0.990	0.020	0.898	0.015	1.157	0.043	1.194	0.008
CAG	Q	1.010	0.020	1.102	0.015	0.843	0.043	0.806	0.008
AAT	N	0.741	0.061	0.795	0.022	0.644	0.018	0.652	0.020
AAC	N	1.259	0.061	1.205	0.022	1.356	0.018	1.348	0.020
AAA	K	0.813	0.014	0.776	0.005	0.870	0.023	0.873	0.007
AAG	K	1.187	0.014	1.224	0.005	1.130	0.023	1.127	0.007
GAT	D	0.809	0.021	0.817	0.019	0.951	0.034	0.935	0.014
GAC	D	1.191	0.021	1.183	0.019	1.049	0.034	1.065	0.014
GAA	E	0.956	0.014	0.880	0.009	0.930	0.014	0.928	0.006
GAG	E	1.044	0.014	1.120	0.009	1.070	0.014	1.072	0.006
TGT	C	1.075	0.026	1.020	0.085	0.923	0.067	0.888	0.026
TGC	C	0.925	0.026	0.980	0.085	1.077	0.067	1.112	0.026
CGT	R	0.387	0.012	0.409	0.022	0.447	0.058	0.472	0.013
CGC	R	0.671	0.014	0.664	0.014	0.601	0.069	0.572	0.016
CGA	R	0.319	0.031	0.338	0.031	0.230	0.020	0.218	0.010
CGG	R	0.530	0.047	0.523	0.026	0.554	0.039	0.576	0.015
AGA	R	2.581	0.030	2.448	0.013	2.393	0.020	2.398	0.013
AGG	R	1.513	0.031	1.619	0.014	1.775	0.037	1.764	0.014
GGT	G	0.524	0.028	0.550	0.008	0.513	0.016	0.520	0.012
GGC	G	0.658	0.019	0.636	0.024	0.692	0.012	0.691	0.011
GGA	G	1.958	0.031	1.915	0.014	1.732	0.024	1.735	0.011
GGG	G	0.860	0.037	0.898	0.003	1.063	0.028	1.054	0.012

Table 2. Pairwise distance of RSCUs, based on root-mean-square deviation (RMSD), between Zika viruses from different locations.

	Asia	Oceania	Americas	N-America	S-America	Europe	E-Africa	W-Africa
Asia		0.0179	0.0178	0.0184	0.0179	0.0170	0.1110	0.1310
Oceania			0.0116	0.0131	0.0105	0.0151	0.1132	0.1349
Americas				0.0030	0.0053	0.0083	0.1143	0.1360
N-America					0.0083	0.0082	0.1137	0.1353
S-America						0.0107	0.1154	0.1373
Europe							0.1146	0.1361
E-Africa								0.0762
W-Africa								

Figure 3. Correlation of RSCUs in Asian, American, Oceanic and European Zika viruses.



RESULTS

2. The results support the African/Asian hypothesis [2] as the origin of Zika virus.

Figure 4. Two hypotheses have been suggested to explain the origin of Zika virus.

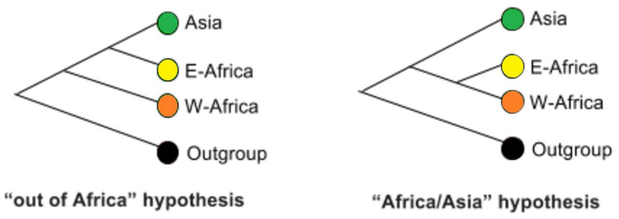
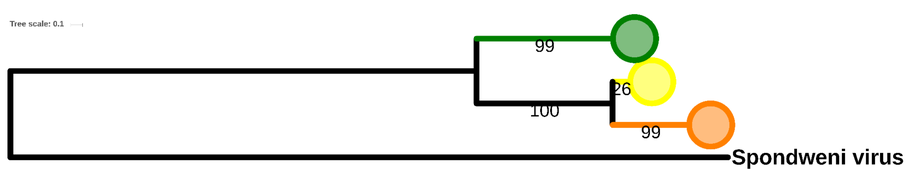


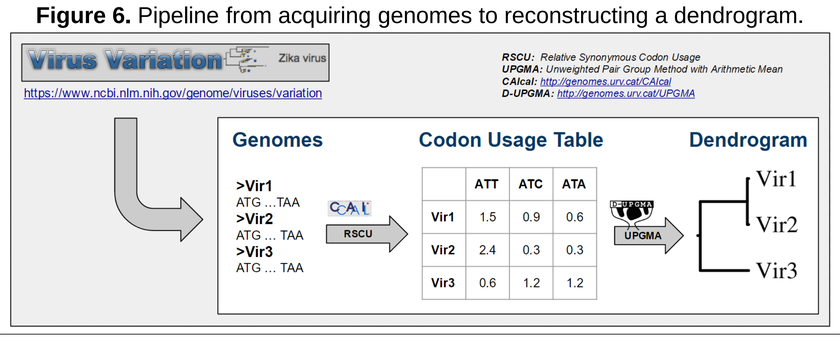
Figure 5. Result tree calculated with Pearson correlations of RSCUs of Asian, East and West African genotypes (100 bootstraps).



MATERIALS AND METHODS

The Zika virus genomes were obtained from NCBI Virus Variation database, and their RSCU was calculated with CAIcal. This data then was analyzed with D-UPGMA to build a dendrogram.

Figure 6. Pipeline from acquiring genomes to reconstructing a dendrogram.



CONCLUSIONS AND FUTURE DIRECTIONS

Conclusions:

- A fast and reliable method to track the phylodynamics of Zika virus based on the relative use of synonymous codons.
- Confirm the Asian origin of the recent Zika outbreak in America.
- Support for the African/Asian hypothesis of the origin of the virus.

Future directions:

- Track the adaptation of the virus to the human host.
- Use in surveillance and risk assessment of future outbreaks.

$$RSCU_{ij} = \frac{x_{ij}}{\frac{i}{n_i}\sum_{j=1}^{n_i} x_{ij}} \qquad RMSD = \sqrt{\frac{\sum(x_i - y_i)^2}{w}}$$

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$

References

1. Zanluca, C. et al. *First report of autochthonous transmission of Zika virus in Brazil*. Mem Inst Oswaldo Cruz Rio Janeiro 110, 569–572 (2015).

2. de Bernardi Schneider, A. et al. *Molecular evolution of Zika virus as it crossed the Pacific to the Americas*. Cladistics 33, 1–20 (2017).