Mutations of SARS-CoV-2 nsp14 exhibit strong association with increased genome-wide mutation load

Doğa Eskier[1, 2, *], Aslı Suner[3, *], Yavuz Oktay[1, 2, 4, †], Gökhan Karakülah[1, 2, †]

[1]Izmir Biomedicine and Genome Center (IBG), Izmir, Turkey

[2]Izmir International Biomedicine and Genome Institute, Dokuz Eylül University, Izmir, Turkey

[3]Department of Biostatistics and Medical Informatics, Faculty of Medicine, Ege University, Izmir, Turkey

[4]Faculty of Medicine, Department of Medical Biology, Dokuz Eylül University, Izmir, Turkey

[*]These authors contributed equally to the work presented here and should therefore be regarded as first authors.

[†]Co-corresponding authors:

Gökhan Karakülah, E-mail: gokhan.karakulah@deu.edu.tr

Yavuz Oktay, E-mail: yavuz.oktay@ibg.edu.tr

Abstract

SARS-CoV-2 is a betacoronavirus responsible for human cases of COVID-19, a pandemic with global impact that first emerged in late 2019. Since then, the viral genome has shown considerable variance as the disease spread across the world, in part due to the zoonotic origins of the virus and the human host adaptation process. As a virus with an RNA genome that codes for its own genomic replication proteins, mutations in these proteins can significantly impact the variance rate of the genome, affecting both the survival and infection rate of the virus, and attempts at combating the disease. In this study, we analyzed the mutation densities of viral isolates carrying frequently observed mutations for four proteins in the RNA synthesis complex over time in comparison to wildtype isolates. Our observations suggest mutations in nsp14, an error-correcting exonuclease protein, have the strongest association with increased mutation load in both regions without selective pressure and across the genome, compared to nsp7, 8, and 12, which form the core polymerase complex. We propose nsp14 as a priority research target for understanding genomic variance rate in SARS-CoV-2 isolates, and nsp14 mutations as potential predictors for high mutability strains.

Keywords: SARS-CoV-2, COVID-19, nsp14, 3'-5' exonuclease, RNA-dependent RNA polymerase, RdRp, mutation rate

Introduction

COVID-19 is an ongoing global pandemic characterized by long-term respiratory system damage in patients, and caused by the SARS-CoV-2 betacoronavirus. It is likely of zoonotic origin, but capable of human-to-human transmission, and since the first observed cases in the Wuhan province of China (Chan et al., 2020; Riou & Althaus, 2020), it has infected over 14 million people, with 612054 recorded deaths. In addition to its immediate effects on the respiratory

system, its long term effects are still being researched, including symptoms such as neuroinvasion (Li, Bai & Hashikawa, 2020; Wu et al., 2020), cardiovascular complications (Kochi et al., 2020; Zhu et al., 2020), and gastrointestinal and liver damage (Lee, Huo & Huang, 2020; Xu et al., 2020). Due to its high transmissibility, and capacity for asymptomatic transmission (Wong et al., 2020), study of COVID-19 and its underlying pathogen remain a high priority. As a result, the high amount of frequently updated data on viral genomes on databases such as GISAID (Elbe & Buckland-Merrett, 2017) and NextStrain (Hadfield et al., 2018) provides researchers with invaluable resources to track the evolution of the virus as it spreads across the world.

SARS-CoV-2 has a linear, single-stranded RNA genome, and does not depend on host proteins for genomic replication, instead using an RNA synthesis complex formed from nonstructural proteins (nsp) coded by its own genome. Four of the key proteins involved in the complex are nsp7, nsp8, nsp12, and nsp14, all of which are formed from cleavage of the polyprotein Orf1ab into mature peptides. Nsp12, also known as RdRp (RNA-dependent RNA polymerase), is responsible for synthesizing new strands of RNA using the viral genome as a template. Nsp7 and nsp8 act as essential co-factors for the polymerase unit, together creating the core polymerase complex (Kirchdoerfer & Ward, 2019; Peng et al., 2020), while nsp14 is an exonuclease which provides error-correcting capability to the RNA synthesis complex (Subissi et al., 2014; Ma et al., 2015; Romano et al., 2020). Owing to their role in maintaining replication fidelity and genome sequences, these proteins are key targets of study in understanding the mutation accumulation and adaptive evolution of the virus (Peng et al., 2020).

In our previous study, we examined the top 10 most frequent mutations in the SARS-CoV-2 nsp12, and identified that four of them are associated with an increase in mutation density in two genes, the membrane glycoprotein (M) and the envelope glycoprotein (E) (the combination of which is hereafter referred to as MoE, as we previously described), which are not under selective

pressure, and mutations in these genes are potential markers of reduced replication fidelity (Eskier et al., 2020). In this study, we follow up on our previous findings and analyze the mutations in nsps 7, 8, and 14, in addition to nsp12, to identify whether the mutations are associated with a nonselective increase in mutation load or not. We then examine whole genome mutation densities in mutant isolates in comparison to wildtype isolates using linear regression models, in order to understand whether the mutations are associated with potential functional impact. Our findings indicate that mutations in nsp14 are most likely to be predictors of accelerated mutation load increase.

Materials and Methods

Genome sequence filtering, retrieval, and preprocessing

As previously described (Eskier et al., 2020), isolate genomes obtained from the GISAID EpiCoV database (date of accession: 17 June, 2020) were filtered to remove low-coverage or incomplete genomes, aligned against the reference genomic sequence for SARS-CoV-2, and processed to identify any SNVs present in the isolates and their impact on peptide sequences, if any, using the MAFFT, snp-sites, bcftools, and ANNOVAR suite of software. Following the alignment and variant annotation, isolates with incomplete sequencing location or date data were further removed from the pool, and the 5' UTR (bases 1-265) and the 100 nucleotides at the 3' end of the genome were masked due to their gap-heavy and low-quality nature. Following the filters, 29,600 genomes were used for the analyses.

Mutation density calculation

Variants were categorized as synonymous and nonsynonymous following annotation by ANNOVAR, with intergenic or terminal mutations being considered synonymous. Gene mutation densities were calculated separately for synonymous and nonsynonymous mutations, as well as

the total of SNVs, for each isolate, using a non-reference nucleotides per kilobase of region metric. Mutation densities were calculated for the combined membrane glycoprotein (M) and envelope glycoprotein (E) genes (MoE), the surface glycoprotein gene (S), and the whole genome.

Statistical Analysis

Descriptive statistics for continuous variable days were calculated with mean, standard deviation, median, and interquartile range. Kolmogorov–Smirnov test was used to check the normality assumption of the continuous variables. In cases of non-normally distributed data, the Wilcoxon rank-sum (Mann-Whitney U) test was performed to determine whether the difference between the two MoE status groups was statistically significant. The Fisher's exact test and the Pearson chi-square test were used for the analysis of categorical variables. The univariate logistic regression method was utilized to assess the mutations associated with MoE status in single variables, and then multiple logistic regression method was performed. The final multiple logistic regression model was executed with the backward stepwise method. The relationship between mutation density and time in isolates with mutations of interest, as well as in the group comprising all isolates, was examined via non-polynomial linear regression model and Spearman's rank correlation. A p-value of less than 0.05 was considered statistically significant. All statistical analyses were performed using IBM SPSS version 25.0 (Chicago, IL, USA).

Results and Discussion

Increases in the mutation load of SARS-CoV-2 are unevenly distributed across its genome

To identify the trends in SARS-CoV-2 mutation load over time, we calculated the average mutation density per day for all isolates for whole genome, S gene, and MoE regions, capping outliers at the 95_{th} and 5_{th} percentile values to minimize the potential effects of sequencing errors

(Fig. 1). Our results show that both at the genome level and the S gene, a very strong positive correlation between average mutation density and time. In comparison, MoE has a weak positive correlation, with a wider spread of mean density in early and late periods compared to the genome and the S gene. This is consistent with reduced selective pressure on the M and E genes, as has previously been described (Dilucca et al., 2020). The top nonsynonymous mutation is 23403A>G (in 22,271 isolates), responsible for the D614G substitution in the spike protein, followed by the 14408C>T mutation (in 22,226 isolates) in the nsp12 region of the Orf1ab gene, causing P323L substitution in the RdRp protein, and the 28144C>T mutation (in 3,081 isolates), responsible for the L84S substitution in the Orf8 protein. The most common synonymous mutation is the 8782C>T mutation (in 3,047 isolates), and is found on the nsp4 coding region of the Orf1ab gene. For the S gene, the most frequent synonymous mutation is the 23731C>T mutation (in 622 isolates), and the second most common nonsynonymous mutation, after the aforementioned D614G mutation, is 25350C>T (in 215 isolates), responsible for the P1263L substitution. For MoE, the most common synonymous and nonsynonymous mutations are 26735C>T (in 341 isolates) and 27046C>T (in 530 isolates), respectively, both of which are found in the M gene, and the latter of which causes T175M amino acid substitution. Other than the D614G mutation, all of the mentioned mutations are C>T substitutions, the prevalence of which in T- or A-rich regions of the SARS-CoV-2 genome have been previously documented (Simmonds, 2020).

Mutations in RNA synthesis complex proteins are associated with higher mutation load

After identifying the increase in mutation load over time, which was more prominent in genes with high functional impact (S, Orf1ab) compared to other structural genes (M, E, N), we sought to examine possible associations of variants in proteins involved in SARS-CoV-2 genome

replication with the increase. We first identified the five most frequently observed mutations for nsps 7, 8, 12 (also known as RdRp) and 14, four of the proteins cleaved from the Orf1ab polyprotein and are involved in the RNA polymerization, followed by analyzing the association of each mutation with the presence of MoE mutations (hereafter referred to as MoE status) using the chi-square test. 12 out of the 20 mutations were found to have a significant association with MoE status (p-value < 0.05) (Table 1). Compared to our previous findings on the top 10 nsp12 mutations (Eskier et al. 2020), which was based on an analysis of 11,208 samples as of 5 May 2020, 13536C>T and 13862C>T have increased in rank of appearance, from 6th and 7th to 4th and 5th, respectively, and decreased in p-value to show statistically significant associations. In addition, the 13730C>T mutation have increased in rank of appearance from 4th to 3rd. Out of the other nsps tested, nsp14 was found to have four significant mutations, while nsp7 had two and nsp8 had one.

Effects of geographical location on MoE status

In addition to time and genotype, we also examined the potential association between the location of isolates and MoE status as a possible confounding factor. We first examined whether there is a significant association between location, defined here as continent the isolate was originally obtained, and MoE status. Our results indicate that there is a strong association between location and MoE status, with the highest percentage of MoE present isolates in Asia (14.5%), and the percentage ratio in South America (6.5%) (p-value <0.001). In comparison to our previous findings, South America had a dramatic decrease in MoE present isolate percentage, likely as a result of the increased sequencing efforts (from 118 isolates to 416) removing potential sampling biases or localized founder effects. Africa, Asia, and North America had an increase in MoE

present proportion, while Europe, Oceania, and South America showed lowered percentages (Table 2).

After observing the potential confounding effect of location on MoE status, we sought to understand whether a location is more or less likely to predict MoE status, using a logistic regression model (Table 3). Comparing each individual region (1) to the other five (0), we found that Asia, Europe, and North and South America are all possible predictors of MoE status (p-value < 0.05), with Asia and Europe 1.697 and 1.184 times as likely to be MoE present as the other regions, and North and South America 0.589 and 0.650 times as likely, respectively.

Using these findings, we created different logistic regression models to identify which of the 12 mutations are likely to be independent predictors of MoE status (Table 4). In the single variable model, all 12 mutations we previously identified and location were found to be potential predictors (p-value < 0.05). Forming final models including the 12 mutations (Final Model A) and the mutations as well as locations (Final Model B), we observed that the predictor effect of two of the mutations nsp8 12478G>A and nsp14 18998C>T do not appear to be sufficiently independent of the other mutations in Final Model A. After adding the location variable to the Final Model A, location remains a significant predictor, with all five non-reference locations less likely to predict MoE than Asia, the reference location, and nsp12 14805C>T is found to not have a predictor effect independent of location (p-value $= 0.073$). Following Final Model B, nine mutations appear to have a significant association with MoE status, independent of other variables: 11916C>T, 12073C>T, 13536C>T, 13730C>T, 13862C>T, 14408C>T, 18060C>T, 18736T>C, and 18877C>T (p-value < 0.05).

Nsp14 mutations have significant impact on increased genomic mutation density

We then examined the effects of each mutation on genomic mutation density to see whether the relationship between the mutations and MoE status are indicative of a genome-wide trend. Due to selection potentially effecting nonsynonymous mutations differentially, we separated the mutations in the two categories and calculated mutation density separately for each category. Our results show that nsp14 mutations show the most consistent association with mutations between MoE and the whole genome. All three nsp14 mutations (18060C>T, 18736T>C, and 18877C>T) which have a significant association with MoE status also show a similar relationship with genomic mutation density (Fig. 2). 18060C>T (L7L) has the lowest odds ratio for MoE status (Table 4), and while it shows a slower increase in synonymous mutation density compared to wildtype isolates (Fig. 2A), it has a significant impact on faster mutation density increase in nonsynonymous mutations (Fig. 2B). In comparison, 18877C>T (L270L) (Fig. 2C-D) and 18736T>C (F233L) (Fig. 2E-F) both show a high prediction capacity for MoE and an increased mutation density. In comparison, mutations in nsp7 (Supp. Figs. 1-2) and nsp12 (Supp. Figs. 3-6) show much lower impact on altered mutation density increase rate. 12073C>T, an nsp7 mutation, displays high divergence from wildtype isolate patterns; however, its low sample size (n = 16) creates a skewed distribution of isolates across time, complicating any potential inference.

It bears noting that not all mutations continue to be observed up to the date of retrieval. 18060C>T and 18736T>C have been last observed on 7 May and 11 May, 2020, respectively. Whether this is due to sequence sampling or because the mutations confer an evolutionary disadvantage is unknown. 18060C>T and 18877C>T are both synonymous mutations, yet only the latter continues to be present in recent isolates.

Conclusions

Our previous work identified RdRp mutations as contributors to the evolution of the SARS-CoV-2 genome and this study confirmed those findings. Furthermore, we hypothesized that mutations of the other critical components of the viral replication and transcription machinery may have similar effects. Our results implicate nsp14 as a source of increased mutation rate in SARS-CoV-2 genomes. Three of the five most common nsp14 mutations, namely 18060C>T, 18736C>T and 18877C>T are associated with increases in both genome-wide mutational load, as well as MoE status, an alternative indicator of mutational rate and virus evolution. Interestingly all three are located within the ExoN domain, which is responsible for the proofreading activity of nsp14; however, only 18736C>T mutation is non-synonymous (F233L), while 18060C>T and 18877C>T are synonymous mutations and therefore, only after functional studies it will be possible to understand their effects on viral replication processes.

The fate of three nsp14 mutations are also intriguing: Despite being present in the first case detected in the Washington state of the US in mid-January, and detected in 1,585 cases till May 7, 18060C>T mutation has not been detected since then. On the other hand, 18877C>T mutation arising around at the end of January likely in Saudi Arabia and being detected in much less cases (n=893), is still present in many isolates. However, it should be noted that 18877C>T mutation arose within the dominant 23403A>G / 14408C>T lineage, while the other two nsp14 mutations are in different lineages. Therefore, dominance or disappearance of different nsp14 mutations may have less to do with these particular mutations and more with the co-mutations. Yet, we cannot rule out possible effects of these nsp14 mutations on the fitness of SARS-CoV-2.

Additional Information and Declarations

Acknowledgement

The authors would like to thank Mr. Alirıza Arıbaş from Izmir Biomedicine and Genome Center for his technical assistance. The authors also would like to extend their thanks to Izmir Biomedicine and Genome Center (IBG) COVID19 platform IBG-COVID19 for their support in implementing the study and the Scientific and Technological Research Council of Turkey (TUBITAK) for their financial support of IBG-COVID19.

Funding

Yavuz Oktay is supported by the Turkish Academy of Sciences Young Investigator Program (TÜBA-GEBİP). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Turkish Academy of Sciences Young Investigator Program (TÜBA-GEBİP).

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

Doğa Eskier, Aslı Suner, Gökhan Karakülah and Yavuz Oktay conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The data is available at Mendeley: Eskier, Doğa; Suner, Aslı; Oktay, Yavuz; Karakülah, Gökhan (2020), "SARS-CoV-2 GISAID isolates (2020-06-17) genotyping VCF", Mendeley Data, v1. http://dx.doi.org/10.17632/63t5c7xb4c.1

Supplemental Information

Supplemental materials are included with this research.

References

Chan JF-W, Yuan S, Kok K-H, To KK-W, Chu H, Yang J, Xing F, Liu J, Yip CC-Y, Poon RW-S, Tsoi H-W, Lo SK-F, Chan K-H, Poon VK-M, Chan W-M, Ip JD, Cai J-P, Cheng VC-C, Chen H, Hui CK-M, Yuen K-Y. 2020. A familial cluster of pneumonia associated with the 2019 novel coronavirus indicating person-to-person transmission: a study of a family cluster. *Lancet (London, England)* 395:514–523. DOI: 10.1016/S0140-6736(20)30154-9.

Dilucca M, Forcelloni S, Georgakilas AG, Giansanti A, Pavlopoulou A. 2020. Codon Usage and Phenotypic Divergences of SARS-CoV-2 Genes. *Viruses* 12:498. DOI: 10.3390/v12050498.

Elbe S, Buckland-Merrett G. 2017. Data, disease and diplomacy: GISAID's innovative contribution to global health. *Global Challenges* 1:33–46. DOI: 10.1002/gch2.1018.

Eskier D, Karakülah G, Suner A, Oktay Y. 2020. RdRp mutations are associated with SARS-CoV-2 genome evolution. *PeerJ* 8:e9587. DOI: 10.7717/peerj.9587.

Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, Sagulenko P, Bedford T, Neher RA. 2018. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics* 34:4121–4123. DOI: 10.1093/bioinformatics/bty407.

Kirchdoerfer RN, Ward AB. 2019. Structure of the SARS-CoV nsp12 polymerase bound to nsp7 and nsp8 co-factors. *Nature Communications* 10:2342. DOI: 10.1038/s41467-019-10280-3.

Kochi AN, Tagliari AP, Forleo GB, Fassini GM, Tondo C. 2020. Cardiac and arrhythmic complications in patients with COVID-19. *Journal of Cardiovascular Electrophysiology* 31:1003–1008. DOI: 10.1111/jce.14479.

Lee I-C, Huo T-I, Huang Y-H. 2020. Gastrointestinal and liver manifestations in patients with COVID-19. *Journal of the Chinese Medical Association: JCMA* 83:521–523. DOI: 10.1097/JCMA.0000000000000319.

Li Y-C, Bai W-Z, Hashikawa T. 2020. The neuroinvasive potential of SARS-CoV2 may play a role in the respiratory failure of COVID-19 patients. *Journal of Medical Virology* 92:552–555. DOI: 10.1002/jmv.25728.

Ma Y, Wu L, Shaw N, Gao Y, Wang J, Sun Y, Lou Z, Yan L, Zhang R, Rao Z. 2015. Structural basis and functional analysis of the SARS coronavirus nsp14–nsp10 complex. *Proceedings of the National Academy of Sciences of the United States of America* 112:9436–9441. DOI: 10.1073/pnas.1508686112.

Peng Q, Peng R, Yuan B, Zhao J, Wang M, Wang X, Wang Q, Sun Y, Fan Z, Qi J, Gao GF, Shi Y. 2020. Structural and Biochemical Characterization of the nsp12-nsp7-nsp8 Core Polymerase Complex from SARS-CoV-2. *Cell Reports* 31:107774. DOI: 10.1016/j.celrep.2020.107774.

Riou J, Althaus CL. 2020. Pattern of early human-to-human transmission of Wuhan 2019 novel coronavirus (2019-nCoV), December 2019 to January 2020. *Eurosurveillance* 25. DOI: 10.2807/1560-7917.ES.2020.25.4.2000058.

Romano M, Ruggiero A, Squeglia F, Maga G, Berisio R. 2020. A Structural View of SARS-CoV-2 RNA Replication Machinery: RNA Synthesis, Proofreading and Final Capping. *Cells* 9. DOI: 10.3390/cells9051267.

Simmonds P. 2020. Rampant C→U Hypermutation in the Genomes of SARS-CoV-2 and Other Coronaviruses: Causes and Consequences for Their Short- and Long-Term Evolutionary Trajectories. *mSphere* 5. DOI: 10.1128/mSphere.00408-20.

Subissi L, Posthuma CC, Collet A, Zevenhoven-Dobbe JC, Gorbalenya AE, Decroly E, Snijder EJ, Canard B, Imbert I. 2014. One severe acute respiratory syndrome coronavirus protein complex integrates processive RNA polymerase and exonuclease activities. *Proceedings of the National Academy of Sciences* 111:E3900–E3909. DOI: 10.1073/pnas.1323705111.

Wong J, Jamaludin SA, Alikhan MF, Chaw L. 2020. Asymptomatic transmission of SARS-CoV-2 and implications for mass gatherings. *Influenza and Other Respiratory Viruses*. DOI: 10.1111/irv.12767.

Wu Y, Xu X, Chen Z, Duan J, Hashimoto K, Yang L, Liu C, Yang C. 2020. Nervous system involvement after infection with COVID-19 and other coronaviruses. *Brain, Behavior, and Immunity* 87:18–22. DOI: 10.1016/j.bbi.2020.03.031.

Xu L, Liu J, Lu M, Yang D, Zheng X. 2020. Liver injury during highly pathogenic human coronavirus infections. *Liver International: Official Journal of the International Association for the Study of the Liver* 40:998–1004. DOI: 10.1111/liv.14435.

Zhu H, Rhee J-W, Cheng P, Waliany S, Chang A, Witteles RM, Maecker H, Davis MM, Nguyen PK, Wu SM. 2020. Cardiovascular Complications in Patients with COVID-19: Consequences of Viral Toxicities and Host Immune Response. *Current Cardiology Reports* 22:32. DOI: 10.1007/s11886-020-01292-3.

Table 1. Comparisons of MoE and nsp mutations.

NSP	Mutations	Values	MoE Absent		MoE Present		Total		p
			n	%	n	%	n	%	
nsp7	11916C>T S3884L	Absent Present	26326 433	98.4 1.6	2833 8	99.7 0.3	29159 441	98.5 1.5	<0.001*
	12076C>T N3937N	Absent Present	26735 24	99.9 0.1	2837 4	99.9 0.1	29572 28	99.9 0.1	0.339
	11919C>T S3885F	Absent Present	26738 21	99.9 0.1	2840 1	100.0 -	29578 22	99.9 0.1	0.717
	12073C>T D3936D	Absent Present	26750 9	100.0 -	2834 7	99.8 0.2	29584 16	99.9 0.1	<0.001*
	11962C>T L3899L	Absent Present	26746 13	100.0 -	2840 1	100.0 -	29586 14	100.0 -	1.000
nsp8	12478G>A M4071I	Absent Present	26757 2	100.0 -	2750 91	96.8 3.2	29507 93	99.7 0.3	<0.001*
	12550G>A L4095L	Absent Present	26697 62	99.8 0.2	2841 -	100.0 -	29538 62	99.8 0.2	-
	12415C>T N4050N	Absent Present	26725 34	99.9 0.1	2841 -	100.0 -	29566 34	99.9 0.1	-
	12557A>G I4098V	Absent Present	26729 30	99.9 0.1	2841 -	100.0 -	29566 30	99.9 0.1	-
	12400C>T L4045L	Absent Present	26734 25	99.9 0.1	2840 1	100.0 -	29574 26	99.9 0.1	0.508
nsp12	14408C>T P4715L	Absent Present	7498 19261	28.0 72.0	702 2139	24.7 75.3	8200 21400	27.7 72.3	<0.001*
	14805C>T Y4847Y	Absent Present	24397 2362	91.2 8.8	2704 137	95.2 4.8	27101 2499	91.6 8.4	<0.001*
	13730C>T A4489V	Absent Present	26238 521	98.1 1.9	2820 21	99.3 0.7	29058 542	98.2 1.8	<0.001*
	13536C>T Y4424Y	Absent Present	26469 290	98.9 1.1	2823 18	99.4 0.6	29292 308	99.0 1.0	0.025*
	13862C>T T4533I	Absent Present	26535 224	99.2 0.8	2833 8	99.7 0.3	29368 232	99.2 0.8	0.001*
nsp14	18060C>T L5932L	Absent Present	25247 1512	94.3 5.7	2768 73	97.4 2.6	28015 1585	94.6 5.4	0.001*
	18877C>T L6205L	Absent Present	26185 574	97.9 2.1	2522 319	88.8 11.2	28707 893	97.0 3.0	0.001*
	18998C>T A6245V	Absent Present	26454 305	98.9 1.1	2836 5	99.8 0.2	29290 310	99.0 1.0	0.001*
	18736T>C F6158L	Absent Present	26751 8	100.0 -	2613 228	92.0 8.0	29364 236	99.2 0.8	0.001*
	19524C>T L6420L	Absent Present	26530 229	99.1 0.9	2825 16	99.4 0.6	29355 245	99.2 0.8	0.102
	Total		26759	100.0	2840	100.0	29600	100.0	

*p-value<0.05 was statistically significant.

Table 2. Distribution of MoE across geographical locations.

Locations	MoE absent		MoE present		Total		p
	n	%	n	%	n	%	
Asia	2319	85.5	394	14.5	2713	100.0	
Africa	297	90.3	32	9.7	329	100.0	
South America	389	93.5	27	6.5	416	100.0	<0.001*
Europe	14879	89.8	1697	10.2	16576	100.0	
North America	7401	93.4	522	6.6	7923	100.0	
Oceania	1474	89.7	169	10.3	1643	100.0	
Total	26759	90.4	2841	9.6	29600	100.0	

*p-value<0.05 was statistically significant.

Table 3. Logistic regression model of MoE and location on single variables. Each location was represented as itself (1) and others (0).

Locations	p	OR	95% C.I.
Asia	<0.001*	1.697	1.513 to 1.903
Africa	0.937	1.015	0.703 to 1.465
South America	0.032*	0.650	0.439 to 0.963
Europe	<0.001*	1.184	1.095 to 1.281
North America	<0.001*	0.589	0.533 to 0.650
Oceania	0.330	1.085	0.921 to 1.278

OR, Odds-Ratio; C.I.: confidence interval, *p-value<0.05 was statistically significant.

Table 4. Logistic regression model of MoE on single variables and a final model. (Final Model A) Logistic regression model of ten mutations on final model. (Final Model B) Logistic regression model of four mutations and location on final model.

Mutations	Single Variables			Final Model A			Final Model B		
	p	OR	95% C.I.	p	OR	95% C.I.	p	OR	95% C.I.
Nsp7.11916	<0.001*	0.172	0.085 to 0.346	<0.001*	0.180	0.089 to 0.363	0.001*	0.314	0.154 to 0.641
Nsp7.12076	0.403	1.571	0.545 to 4.530	-	-	-	-	-	-
Nsp7.11919	0.433	0.448	0.060 to 3.334	-	-	-	-	-	-
Nsp7.12073	<0.001*	7.341	2.732 to 19.728	<0.001*	8.108	3.009 to 21.847	<0.001*	9.164	3.311 to 25.361
Nsp7.11962	0.756	0.724	0.095 to 5.540	-	-	-	-	-	-
Nsp8.12478	<0.001*	442.707	108.996 to 1798.139	-	-	-	-	-	-
Nsp8.12550	0.997	-	-	-	-	-	-	-	-
Nsp8.12415	0.998	-	-	-	-	-	-	-	-
Nsp8.12557	0.998	-	-	-	-	-	-	-	-
Nsp8.12400	0.388	0.377	0.051 to 2.780	-	-	-	-	-	-
Nsp12.14408	<0.001*	1.186	1.085 to 1.297	<0.001*	1.310	1.144 to 1.500	<0.001*	1.662	1.435 to 1.926
Nsp12.14805	<0.001*	0.523	0.439 to 0.625	0.007*	0.746	0.603 to 0.923	0.073	0.817	0.655 to 1.019
Nsp12.13730	<0.001*	0.375	0.242 to 0.581	0.002*	0.497	0.317 to 0.778	<0.001*	0.393	0.250 to 0.619
Nsp12.13536	0.026*	0.582	0.361 to 0.938	0.044*	0.611	0.379 to 0.987	0.009*	0.528	0.327 to 0.855
Nsp12.13862	0.002*	0.335	0.165 to 0.678	0.004*	0.355	0.175 to 0.720	0.001*	0.293	0.144 to 0.594
Nsp14.18060	<0.001*	0.440	0.347 to 0.559	0.001*	0.625	0.479 to 0.816	0.001*	1.658	1.244 to 2.209
Nsp14.18877	<0.001*	5.770	5.002 to 6.656	<0.001*	5.543	4.793 to 6.409	<0.001*	6.437	5.483 to 7.557
Nsp14.18998	<0.001*	0.153	0.063 to 0.370	-	-	-	-	-	-
Nsp14.18736	<0.001*	291.773	144.002 to 591.182	<0.001*	368.884	180.195 to 755.153	<0.001*	970.884	469.324 to 2008.453
Nsp14.19524	0.104	0.656	0.395 to 1.091	-	-	-	-	-	-
Location	<0.001*	-	-				<0.001*		
Africa	0.019*	0.634	0.434-0.927	-	-	-	0.017*	0.580	0.391-0.860
South America	<0.001*	0.409	0.273-0.612	-	-	-	<0.001*	0.302	0.198-0.461
Europe	<0.001*	0.671	0.597-0.755	-	-	-	<0.001*	0.681	0.591-0.785
North America	<0.001*	0.415	0.361-0.477	-	-	-	<0.001*	0.228	0.192-0.271
Oceania	<0.001*	0.675	0.557-0.817	-	-	-	<0.001*	0.536	0.428-0.670

OR: Odds-Ratio; C.I.: confidence interval; Multiple logistic regression final model was executed on all these statistically

significant variables, included together in the model, and selected with backward stepwise method; *p-

value<0.05 was statistically significant.

Figure Legends

Figure 1. The average mutation density per day for genome, S gene, and M and E genes.

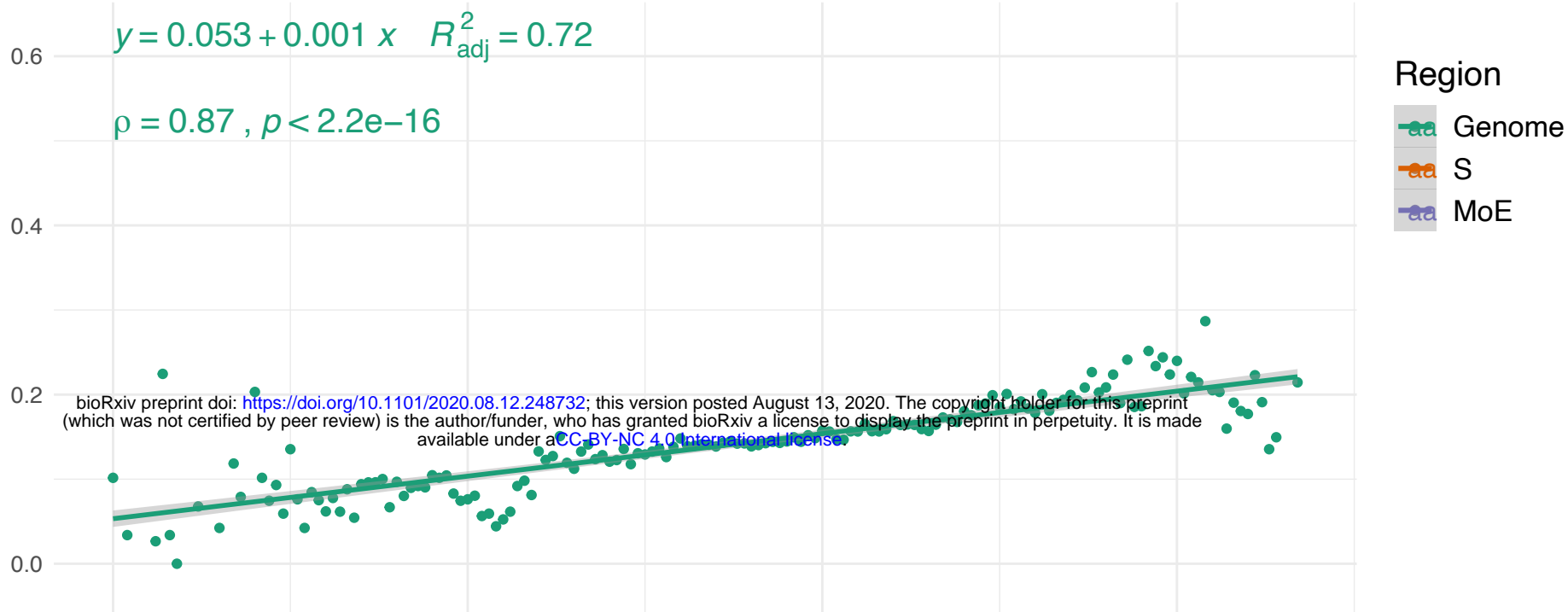Correlation scores are calculated using Spearman rank correlation.

Figure 2. The distribution of synonymous and nonsynonymous mutations in isolates carrying nsp14 mutations compared to wildtype isolates. (A-B) Isolates carrying the synonymous18060C>T mutation (n = 1585). (C-D) Isolates carrying the synonymous 18877C>T mutation (n = 893). (E-F) Isolates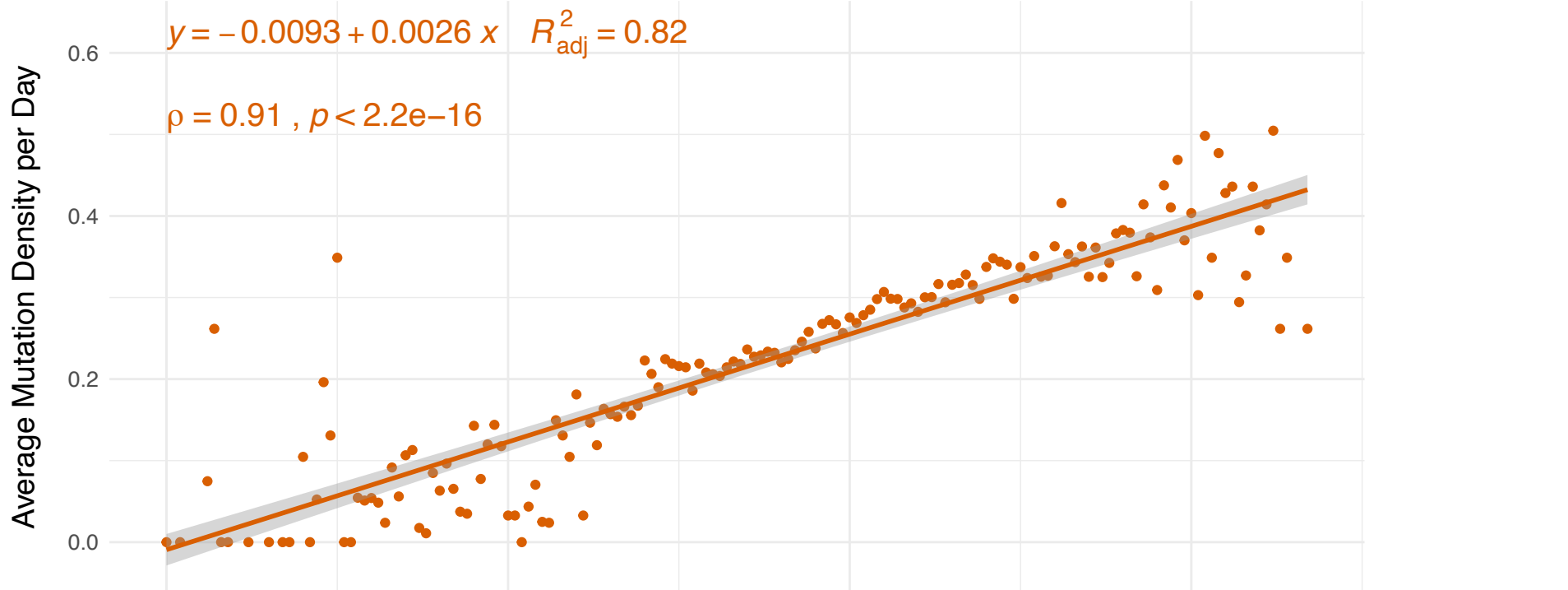 carrying the nonsynonymous 18736C>T mutation (n=236). Wildtype isolates in all graphs carry the reference nucleotide for the nine positions of interest (11916, 12073, 13536, 13730, 13862, 14408, 18060, 18736, 18877) (n = 5910). Correlation scores are calculated using Spearman rank correlation.
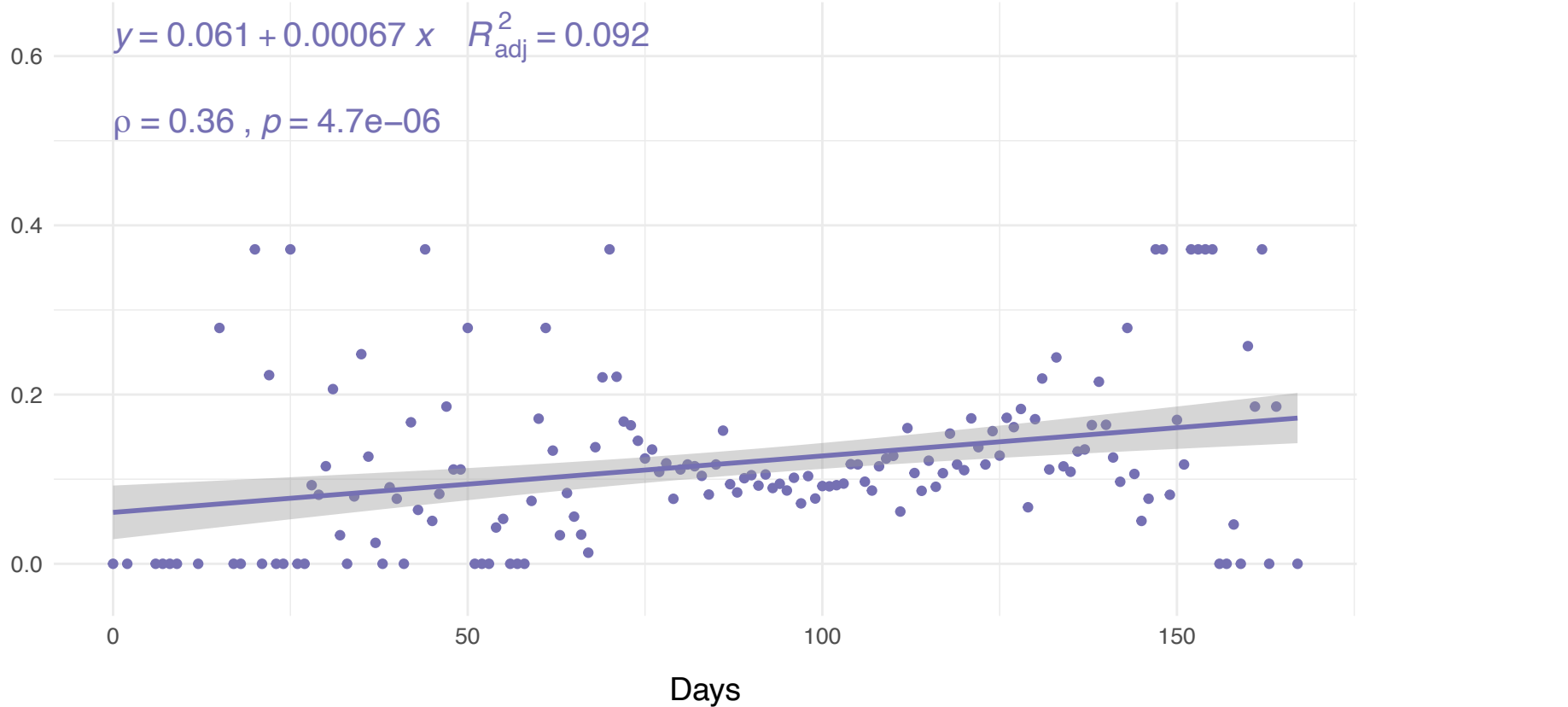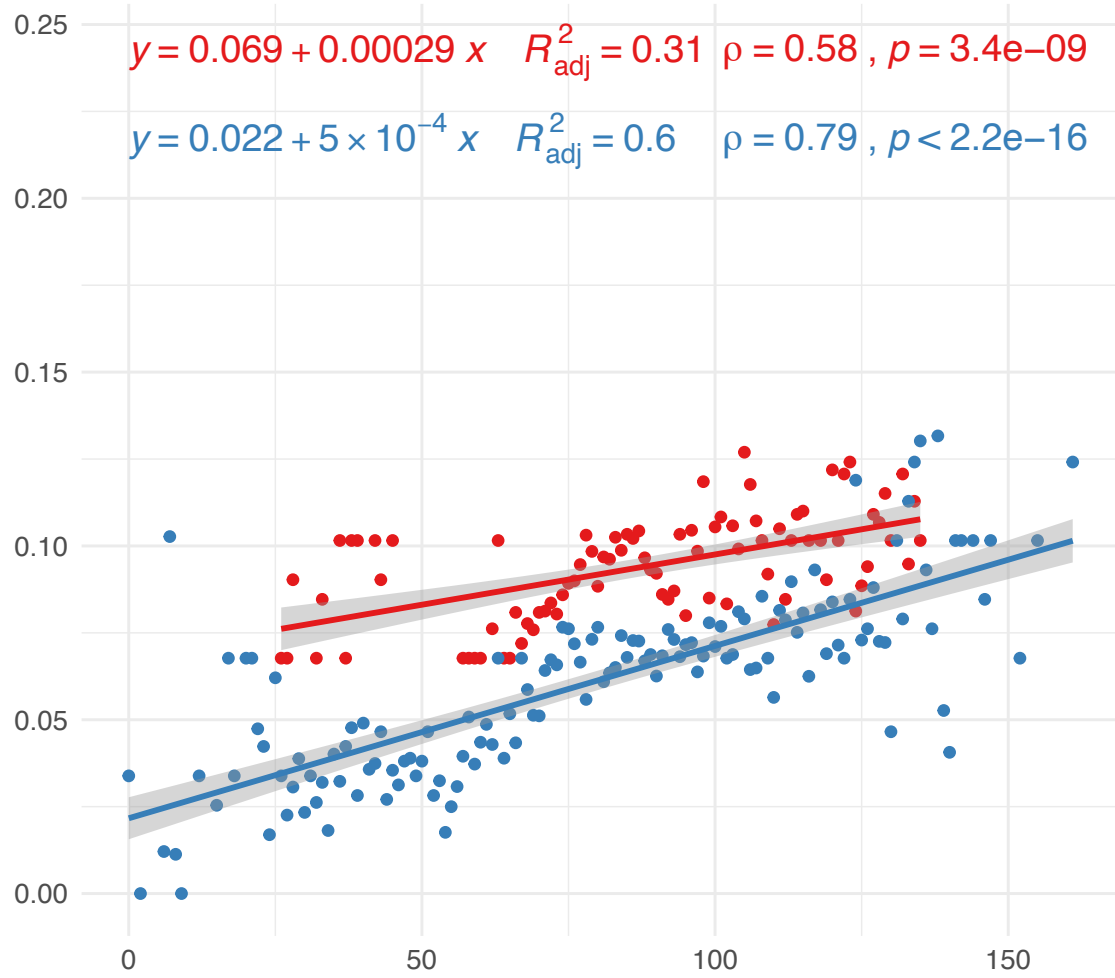
Average Density vs Days – All

A.

$y = 0.053 + 0.001\,x \quad R^2_{adj} = 0.72$

$\rho = 0.87\ ,\ p < 2.2e{-}16$

B.

$y = -0.0093 + 0.0026\,x \quad R^2_{adj} = 0.82$

$\rho = 0.91\ ,\ p < 2.2e{-}16$

C.

$y = 0.061 + 0.00067\,x \quad R^2_{adj} = 0.092$

$\rho = 0.36\ ,\ p = 4.7e{-}06$

Average Mutation Density per Day

Days

Region

Genome

S

MoE

Synonymous

Nonsynonymous

A.

$y = 0.069 + 0.00029\,x$ $R^2_{adj} = 0.31$ $\rho = 0.58$, $p = 3.4e{-}09$

$y = 0.022 + 5 \times 10^{-4}\,x$ $R^2_{adj} = 0.6$ $\rho = 0.79$ $p < 2.2e{-}16$

B.

$y = 0.019 + 0.0012\,x$ $R^2_{adj} = 0.79$ $\rho = 0.89$, $p < 2.2e{-}16$

$y = 0.03 + 0.00055\,x$ $R^2_{ad} = 0.55$ $\rho = 0.77$, $p < 2.2e{-}16$

C.

$y = -0.033 + 0.001\,x$ $R^2_{adj} = 0.59$ $\rho = 0.76$ $p < 2.2e{-}16$

$y = 0.022 + 5 \times 10^{-4}\,x$ $R^2_{adj} = 0.6$ $\rho = 0.79$ $p < 2.2e{-}16$

D.

$y = 0.018 + 0.00083\,x$ $R^2_{adj} = 0.6$ $\rho = 0.82$, $p < 2.2e{-}16$

$y = 0.03 + 0.00055\,x$ $R^2_{adj} = 0.55$ $\rho = 0.77$, $p < 2.2e{-}16$

E.

$y = 0.031 + 0.00072\,x$ $R^2_{adj} = 0.25$ $\rho = 0.55$, $p = 0.00011$

$y = 0.022 + 5 \times 10^{-4}\,x$ $R^2_{adj} = 0.6$ $\rho = 0.79$ $p < 2.2e{-}16$

F.

$y = 0.029 + 0.0019\,x$ $R^2_{adj} = 0.56$ $\rho = 0.76$, $p = 3.3e{-}09$

$y = 0.03 + 0.00055\,x$ $R^2_{adj} = 0.55$ $\rho = 0.77$, $p < 2.2e{-}16$

Genotype

mutant

wildtype

Average Mutation Density per Day

Days