Supplemental Information for:

Investigation of recombination-intense viral groups and their genes in the Earth's virome

Jan P. Meier-Kolthoff[1], Jumpei Uchiyama[2], Hiroko Yahara[3], David Paez-Espino[4], and Koji Yahara[5*]

[1] Department of Bioinformatics, Leibniz Institute DSMZ – German Collection of Microorganisms and Cell Cultures, 38124 Braunschweig, Germany

[2] School of Veterinary Medicine, Azabu University, Sagamihara, Kanagawa, 252-0206, Japan

[3] Department of Cell Signaling, Graduate School of Medical and Dental Sciences, Tokyo Medical and Dental University, Yushima 1-5-45, Bunkyo-ku, Tokyo, 113-8549, Japan

[4] Department of Energy, Joint Genome Institute, Walnut Creek, California 94598, USA

[5] Antimicrobial Resistance Research Center, National Institute of Infectious Diseases, Higashimurayama, Tokyo, 208-0011, Japan

[*]Corresponding author: Koji Yahara (k-yahara@nih.go.jp)

Supplementary Tables

Table S1, S2, and S3 are available as spreadsheets.

Supplementary figure legends

Figure S1. A 16S rRNA gene maximum likelihood tree of the viral groups' host bacterial species.

Figure S2. Distribution of the deviation from the regression line among the 211 viral groups. The dashed vertical line is the cutoff (> 0.009) to define the recombination-intense viral groups.

Figure S3. Mapping reads back to genomes of recombination-intense viral groups. For each viral group, a representative with the longest possible scaffold was used as a reference for the read mapping. Colors of genes are defined at the end according with their predicted function (based on clusters of orthologous groups; COGs).

Figure S4. Phage proteomic tree based on the VICTOR method using a united dataset of comprehensive ICTV reference data and the recombination-intense viral groups. The numbers in circles indicate the assignment of known phages to the ICTV classification at the species (1), genus (3), subfamily (5), and family (7) levels. The remaining numbers indicate the assignment to proteome-based species (2), genus (4), subfamily (6), and family clusters (8) as inferred by the VICTOR method. Leaf labels representing the recombination-intense viral groups are highlighted in orange. The vicinity of these metagenomic samples to actual ICTV phage species, genera, and (sub-)families provides hints regarding their composition. Scale bar indicates interproteomic distances calculated via the distance formula d_4. The tree was rooted at the midpoint[69].

Figure S5. Notable recombination-intense genes deviating from the regression of r_{min} per gene length on nucleotide diversity in each viral group. Viral groups not in Figure 4 are shown. The x-axis and y-axis are the same as in Figure 2 and Figure 4. Pink: notable recombination-intense genes (Figure 5 and Table S2). Green: uncharacterized genes. Gray: others. Recombination breakpoints in a phage tail gene and a portal gene are shown as red vertical bars at the top. The dashed line indicates the linear regression controlling for the number of sequences in each gene.

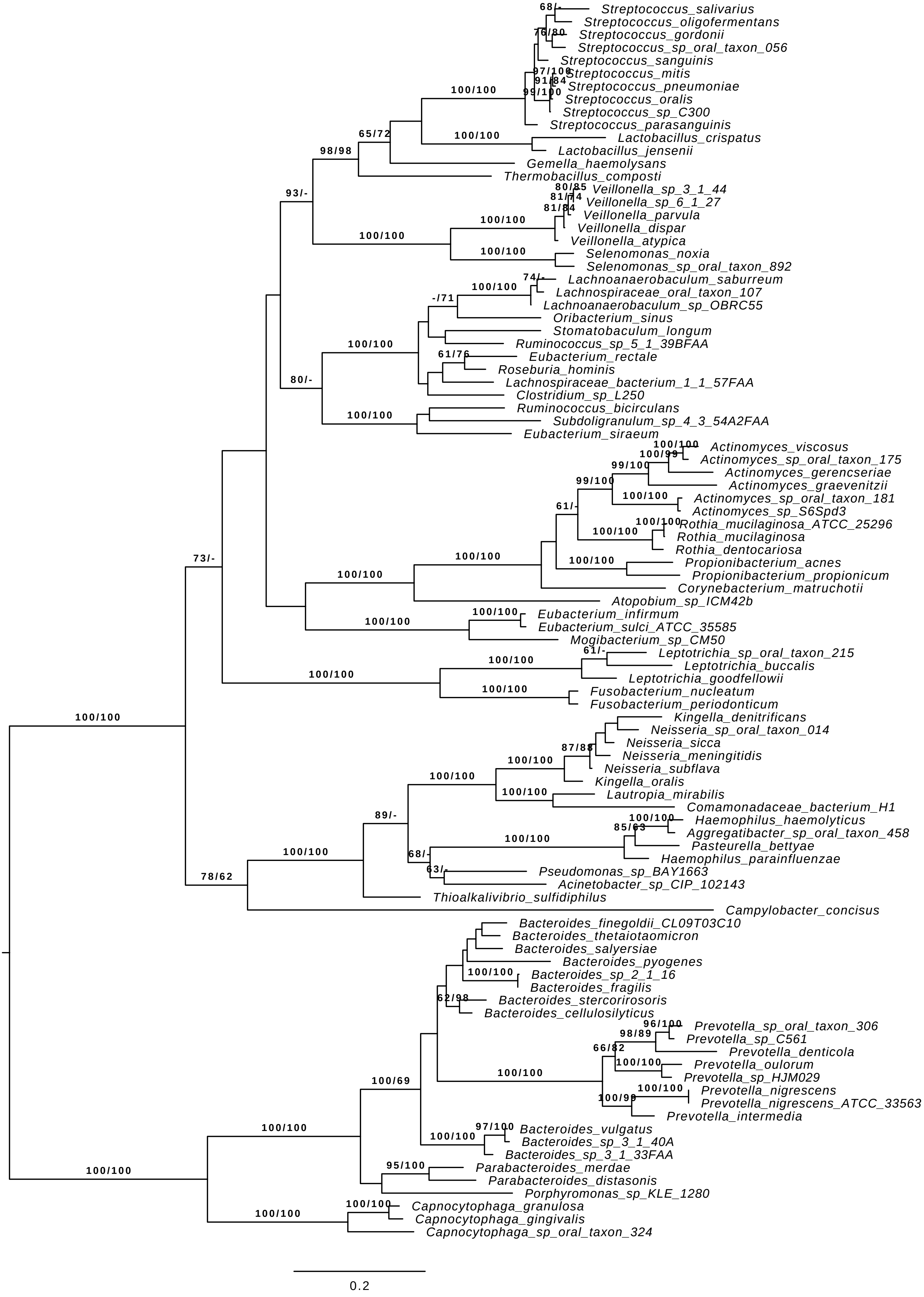
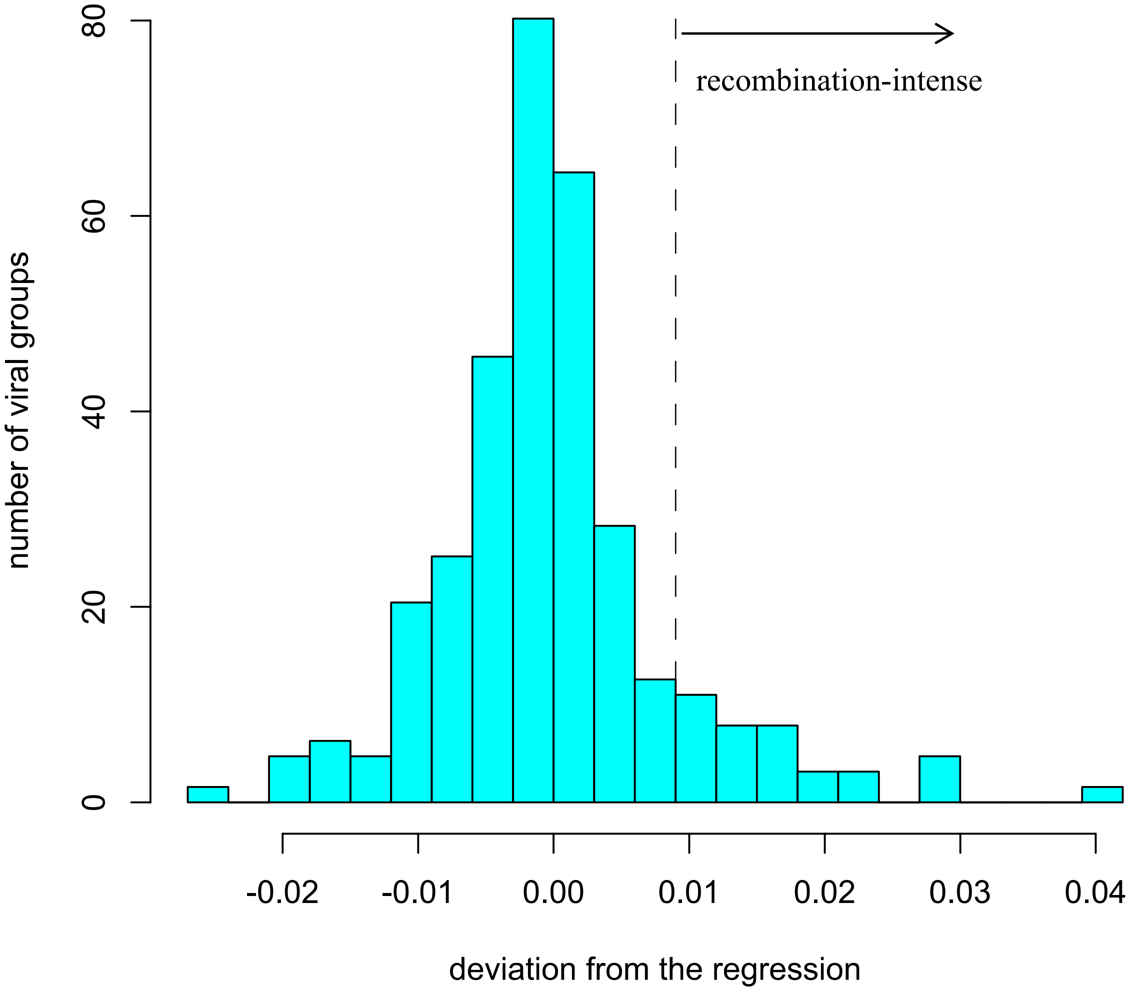
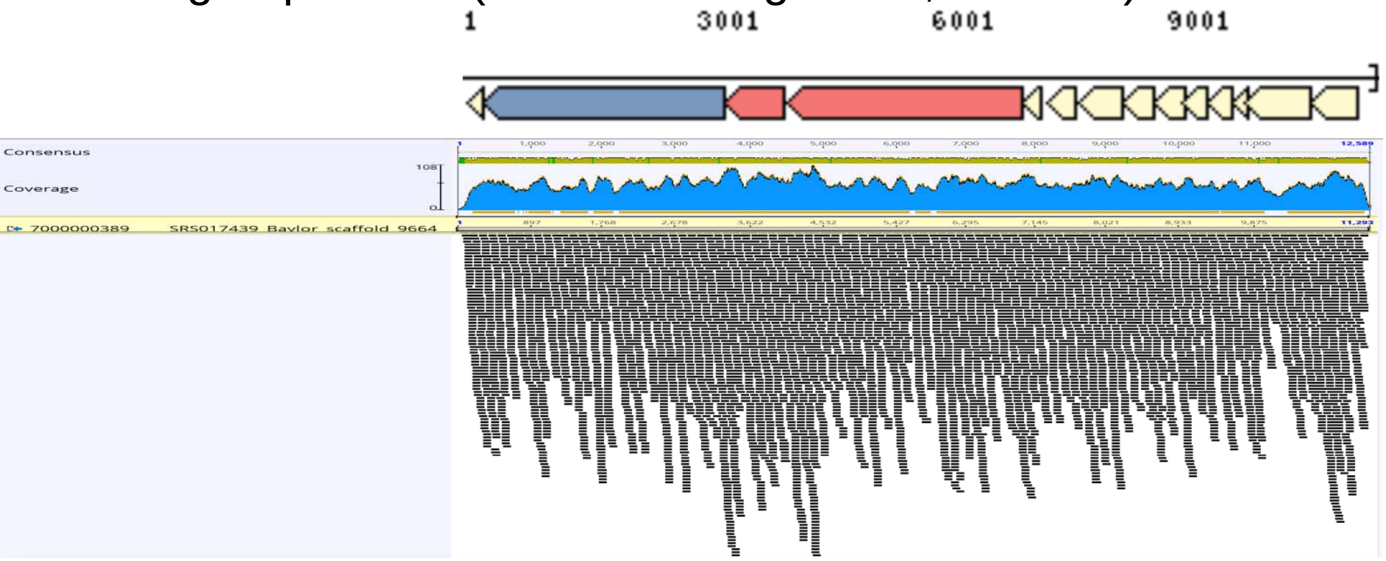

68/-
Streptococcus_salivarius
Streptococcus_oligofermentans
76/80
Streptococcus_gordonii
Streptococcus_sp_oral_taxon_056
Streptococcus_sanguinis
97/100
Streptococcus_mitis
91/84
Streptococcus_pneumoniae
99/100
Streptococcus_oralis
Streptococcus_sp_C300
Streptococcus_parasanguinis
100/100
100/100
Lactobacillus_crispatus
Lactobacillus_jensenii
65/72
98/98
Gemella_haemolysans
Thermobacillus_composti
93/-
80/85
Veillonella_sp_3_1_44
81/74
Veillonella_sp_6_1_27
81/84
Veillonella_parvula
100/100
Veillonella_dispar
Veillonella_atypica
100/100
100/100
Selenomonas_noxia
Selenomonas_sp_oral_taxon_892
74/-
Lachnoanaerobaculum_saburreum
100/100
Lachnospiraceae_oral_taxon_107
-/71
Lachnoanaerobaculum_sp_OBRC55
Oribacterium_sinus
Stomatobaculum_longum
100/100
Ruminococcus_sp_5_1_39BFAA
61/76
Eubacterium_rectale
Roseburia_hominis
80/-
Lachnospiraceae_bacterium_1_1_57FAA
Clostridium_sp_L250
Ruminococcus_bicirculans
100/100
Subdoligranulum_sp_4_3_54A2FAA
Eubacterium_siraeum
100/100
Actinomyces_viscosus
100/99
Actinomyces_sp_oral_taxon_175
99/100
Actinomyces_gerencseriae
99/100
Actinomyces_graevenitzii
100/100
Actinomyces_sp_oral_taxon_181
61/-
Actinomyces_sp_S6Spd3
100/100
Rothia_mucilaginosa_ATCC_25296
100/100
Rothia_mucilaginosa
Rothia_dentocariosa
100/100
100/100
Propionibacterium_acnes
Propionibacterium_propionicum
100/100
Corynebacterium_matruchotii
Atopobium_sp_ICM42b
73/-
100/100
Eubacterium_infirmum
100/100
Eubacterium_sulci_ATCC_35585
Mogibacterium_sp_CM50
61/-
Leptotrichia_sp_oral_taxon_215
100/100
Leptotrichia_buccalis
100/100
Leptotrichia_goodfellowii
100/100
Fusobacterium_nucleatum
100/100
Fusobacterium_periodonticum
Kingella_denitrificans
Neisseria_sp_oral_taxon_014
87/88
Neisseria_sicca
100/100
Neisseria_meningitidis
Neisseria_subflava
100/100
Kingella_oralis
100/100
Lautropia_mirabilis
Comamonadaceae_bacterium_H1
89/-
100/100
Haemophilus_haemolyticus
85/63
Aggregatibacter_sp_oral_taxon_458
100/100
Pasteurella_bettyae
68/-
Haemophilus_parainfluenzae
100/100
63/-
Pseudomonas_sp_BAY1663
78/62
Acinetobacter_sp_CIP_102143
Thioalkalivibrio_sulfidiphilus
Campylobacter_concisus
Bacteroides_finegoldii_CL09T03C10
Bacteroides_thetaiotaomicron
Bacteroides_salyersiae
Bacteroides_pyogenes
100/100
Bacteroides_sp_2_1_16
Bacteroides_fragilis
62/98
Bacteroides_stercorirosoris
Bacteroides_cellulosilyticus
96/100
Prevotella_sp_oral_taxon_306
98/89
Prevotella_sp_C561
66/82
Prevotella_denticola
100/100
Prevotella_oulorum
100/100
Prevotella_sp_HJM029
100/100
Prevotella_nigrescens
100/99
Prevotella_nigrescens_ATCC_33563
100/69
Prevotella_intermedia
97/100
Bacteroides_vulgatus
100/100
100/100
Bacteroides_sp_3_1_40A
Bacteroides_sp_3_1_33FAA
95/100
Parabacteroides_merdae
100/100
Parabacteroides_distasonis
Porphyromonas_sp_KLE_1280
100/100
Capnocytophaga_granulosa
100/100
Capnocytophaga_gingivalis
Capnocytophaga_sp_oral_taxon_324
0.2

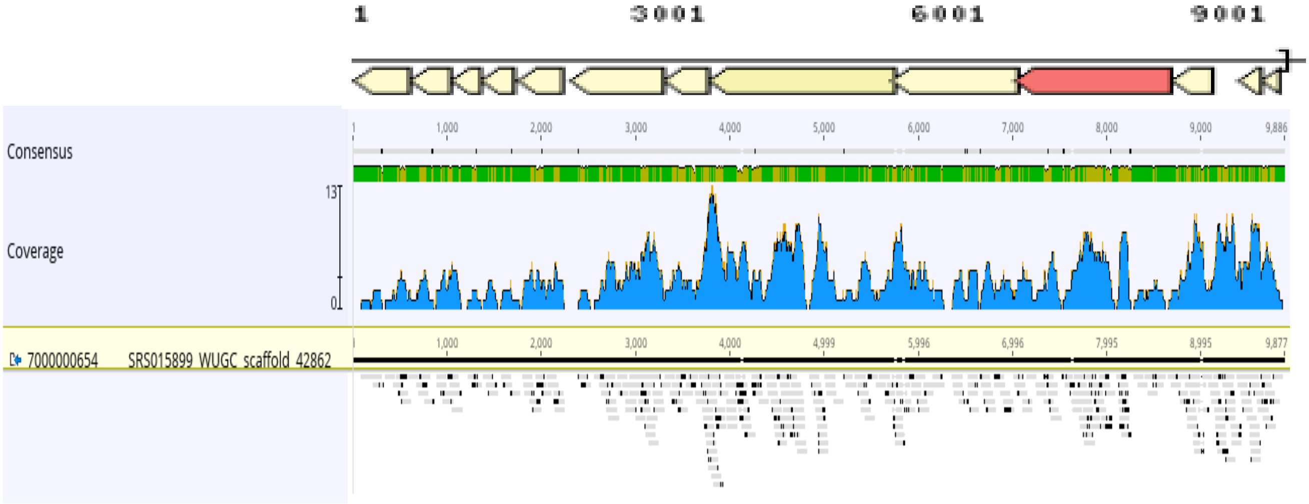
Figure S2


80
60
40
20
0
number of viral groups
recombination-intense
-0.02
-0.01
0.00
0.01
0.02
0.03
0.04
deviation from the regression

viralgroup 2961 (mean coverage 64.1, SD 13.8)



viralgroup 2959 (mean coverage 3.4, SD 2.4)



viralgroup 4460 (mean coverage 7.5, SD 4.2)

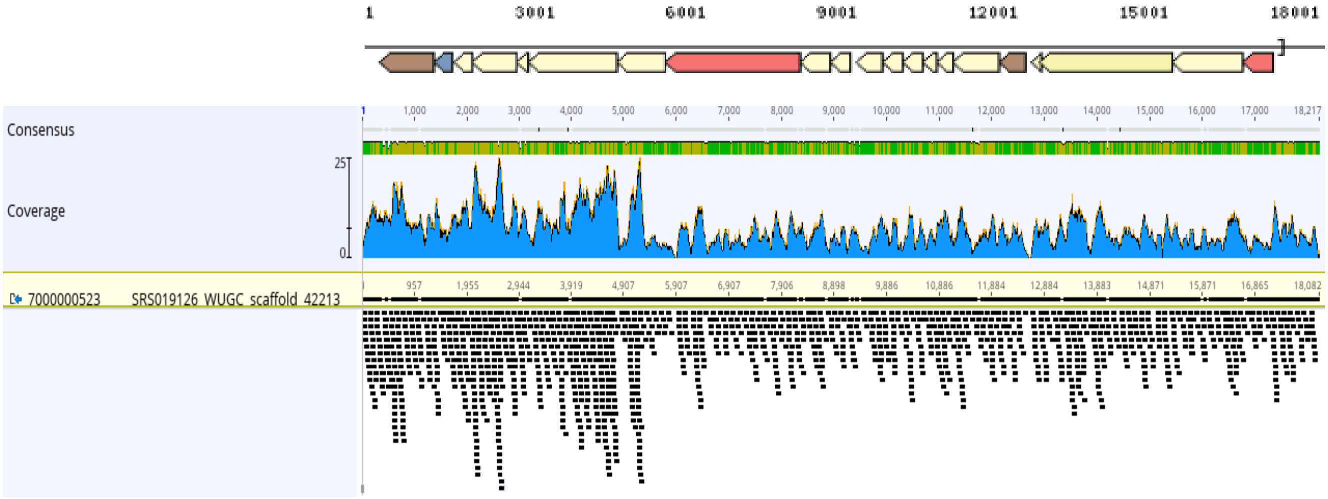
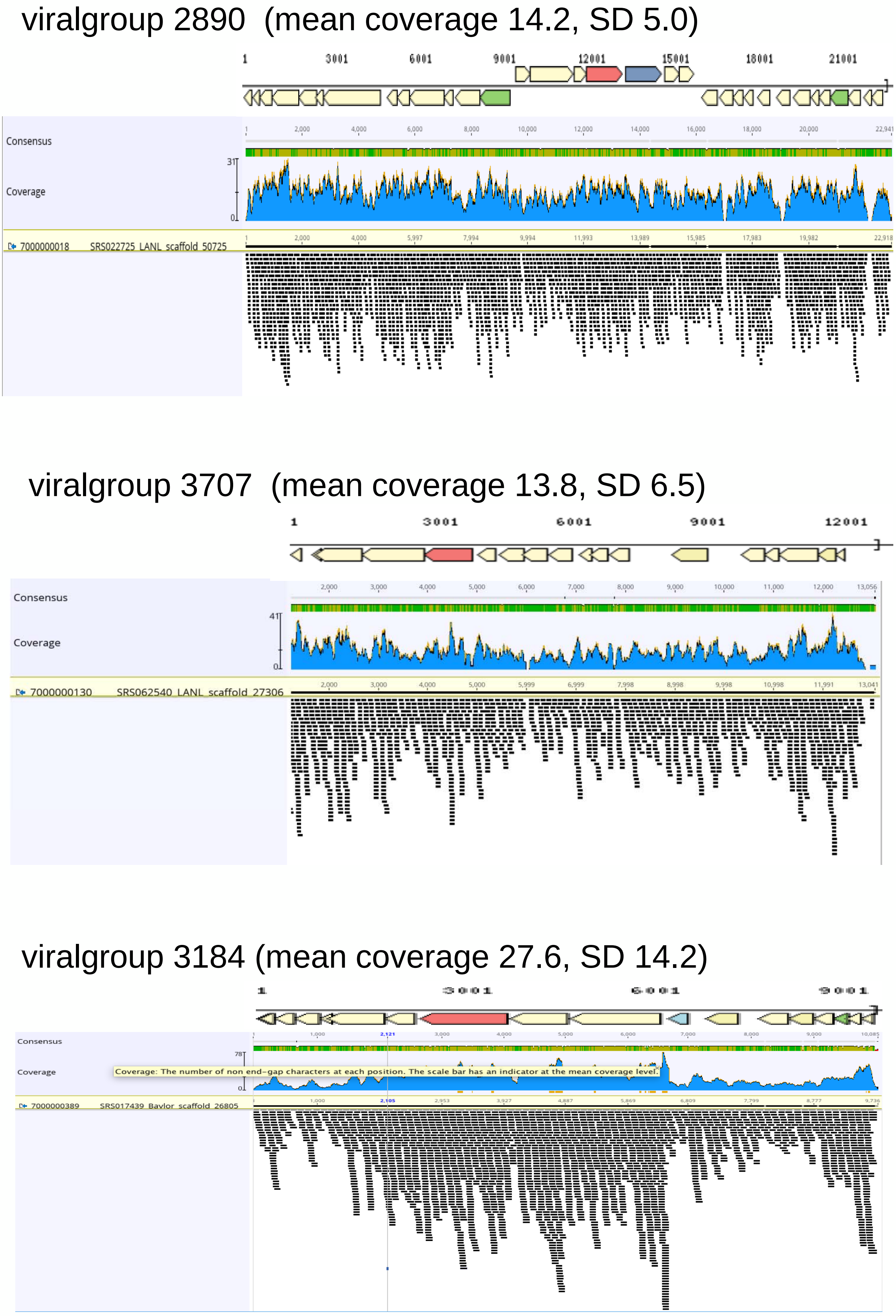


Figure S3



viralgroup 2890 (mean coverage 14.2, SD 5.0)

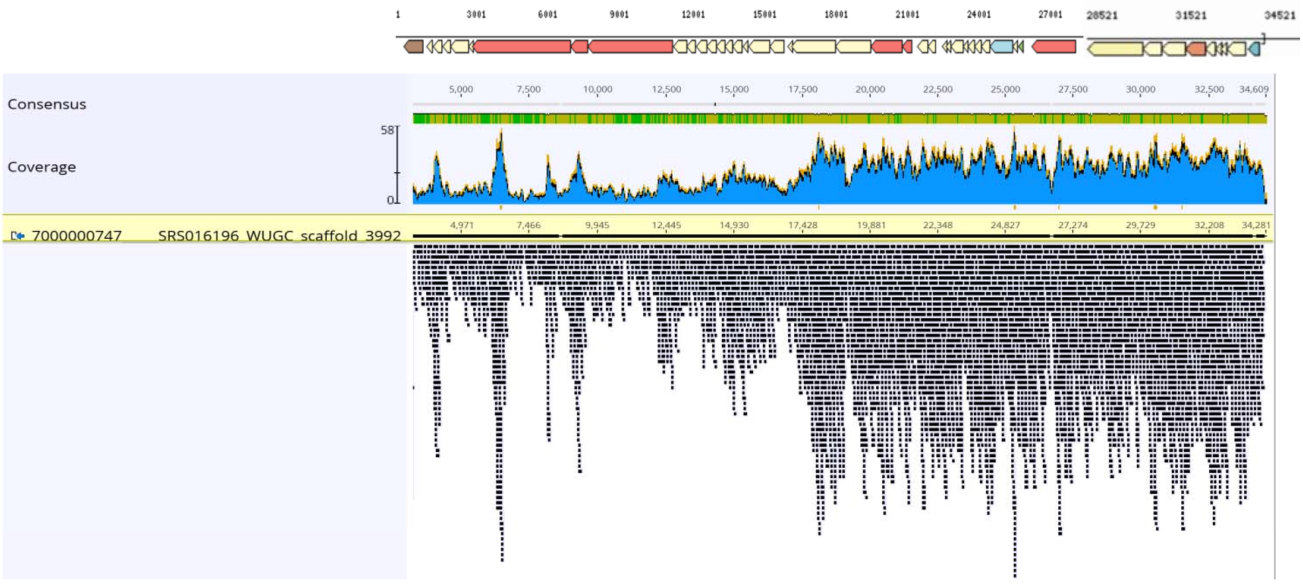
viralgroup 3707 (mean coverage 13.8, SD 6.5)

viralgroup 3184 (mean coverage 27.6, SD 14.2)

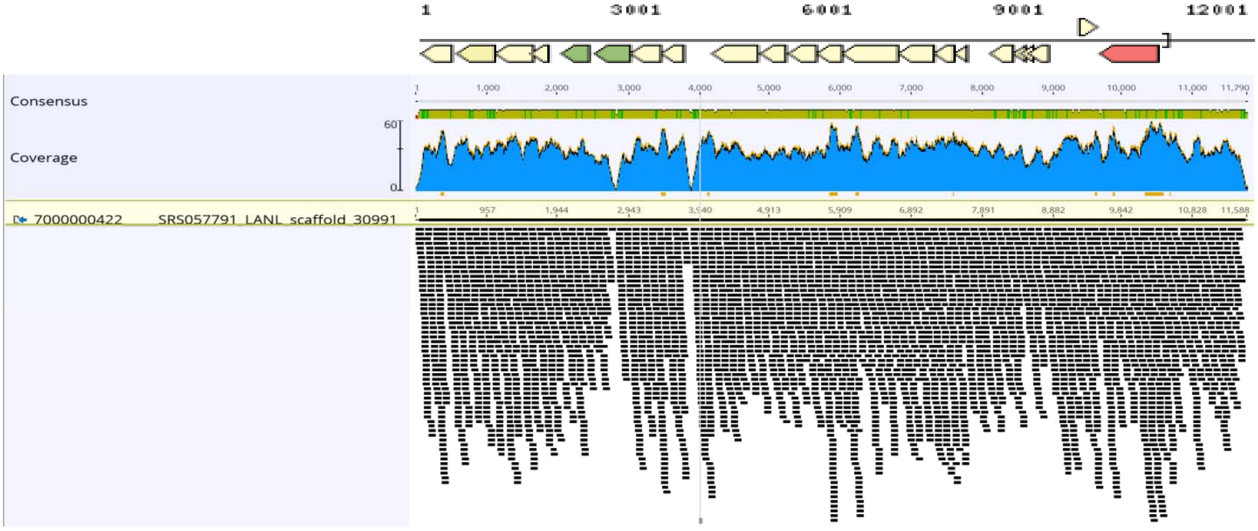
Coverage: The number of non end-gap characters at each position. The scale bar has an indicator at the mean coverage level.

Figure S3

viralgroup 3319 (mean coverage 10.5, SD 4.0)



viralgroup 3776 (mean coverage 35.7, SD 8.5)



viralgroup 2970 (mean coverage 9.1, SD 3.3)

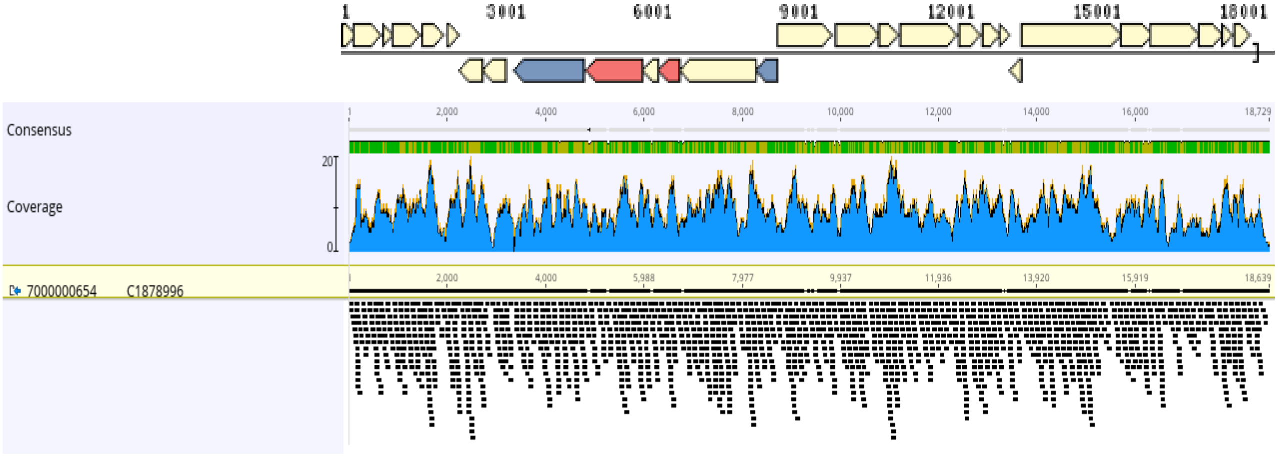
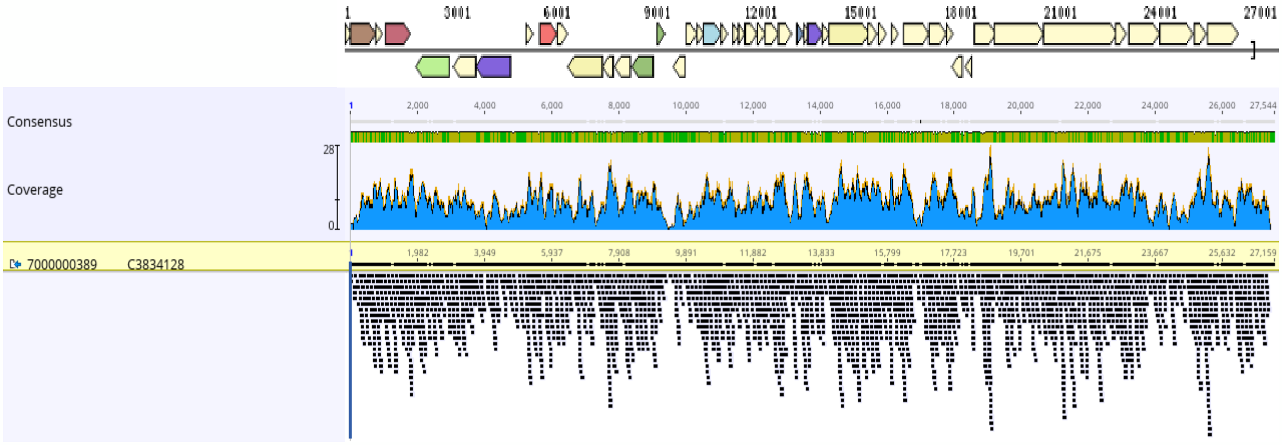
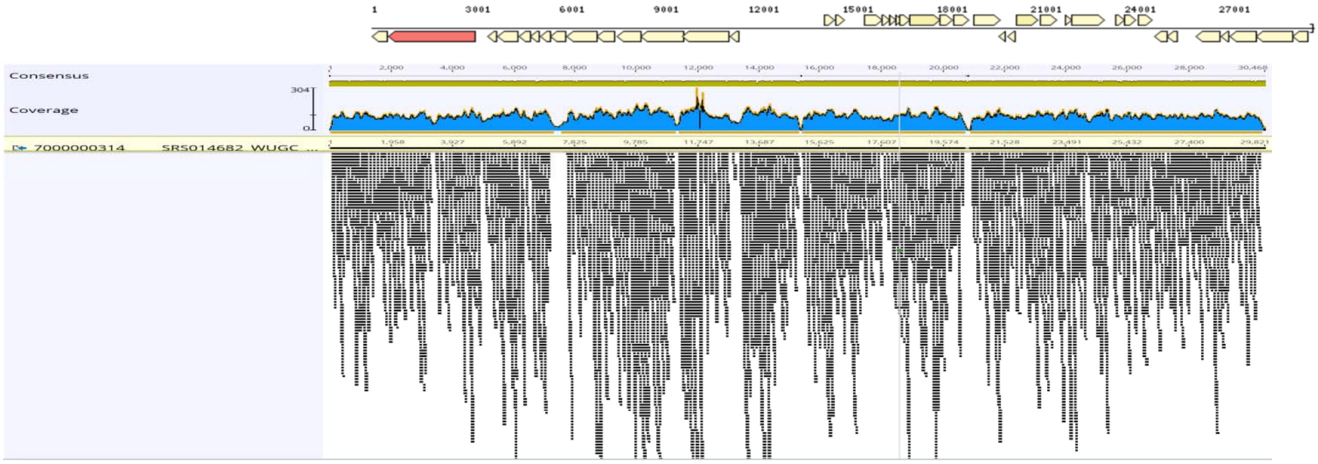


Figure S3

viralgroup 3576 (mean coverage 9.7, SD 4.3)



viralgroup 181 (mean coverage106.9, SD 27.8)



viralgroup 3189 (mean coverage 10.5, SD 4.0)

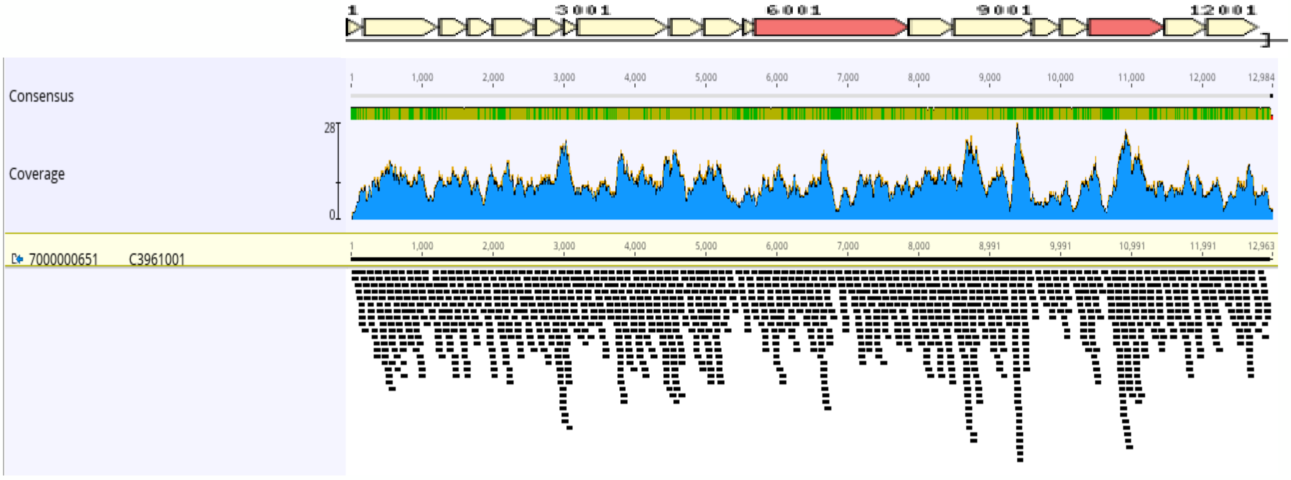
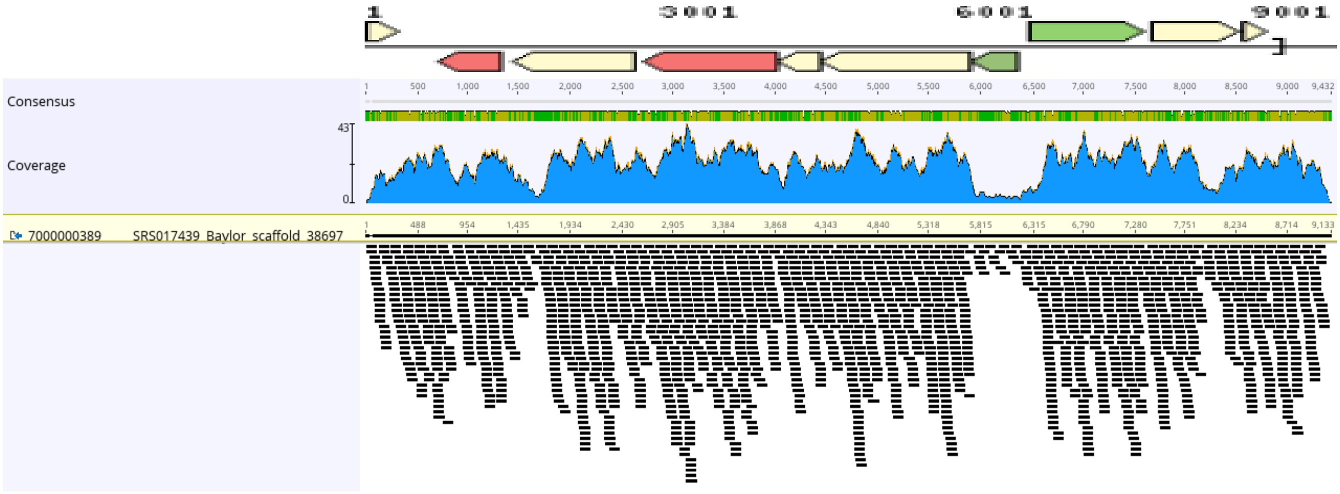
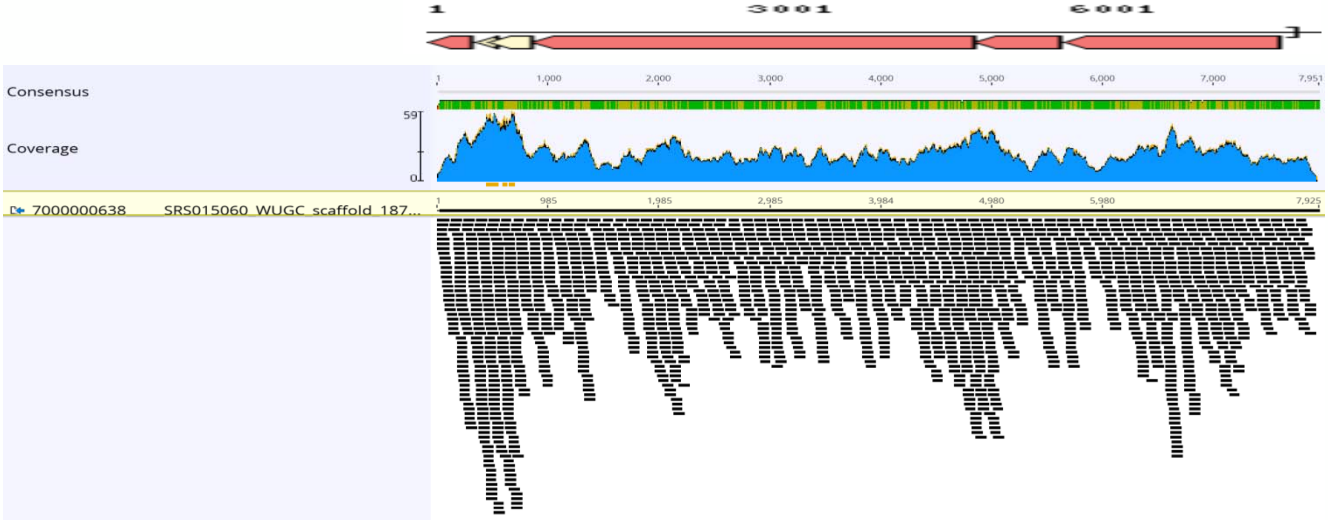


Figure S3

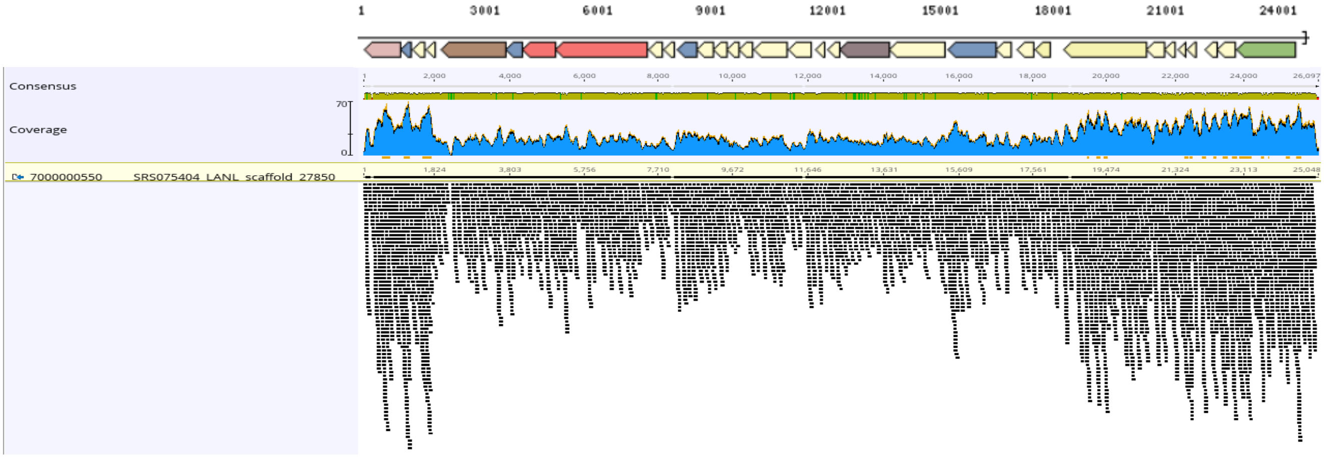
viralgroup 3549 (mean coverage 21.2, SD 8.2)



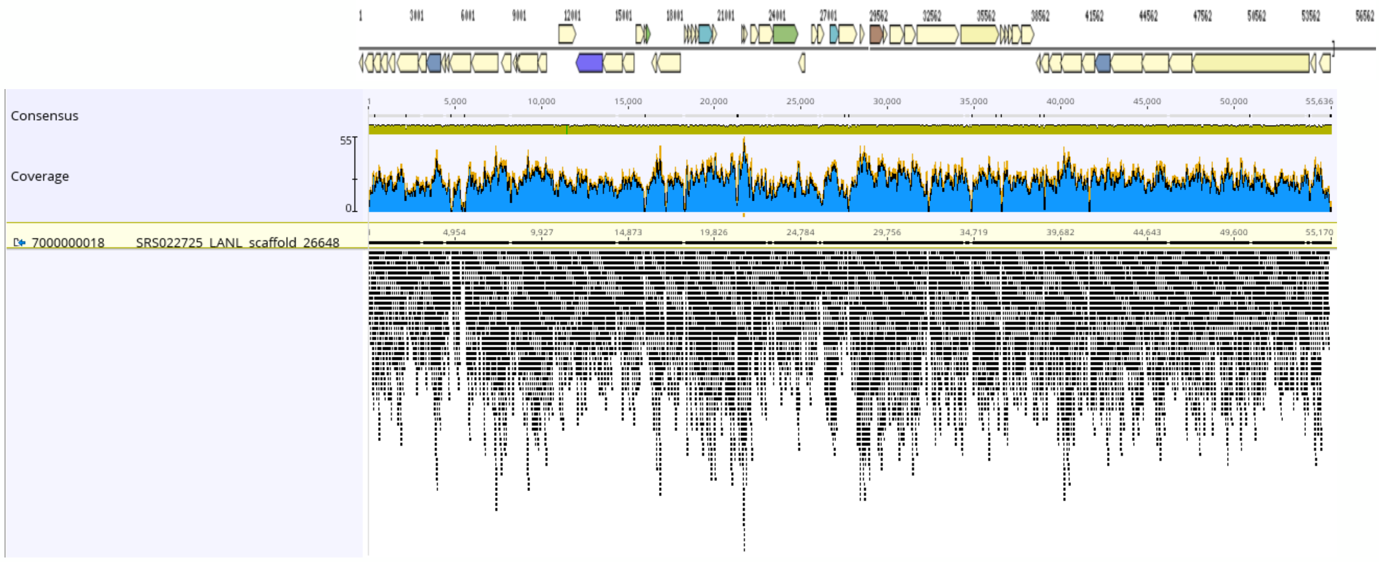
viralgroup 3600 (mean coverage 24.6, SD 9.0)



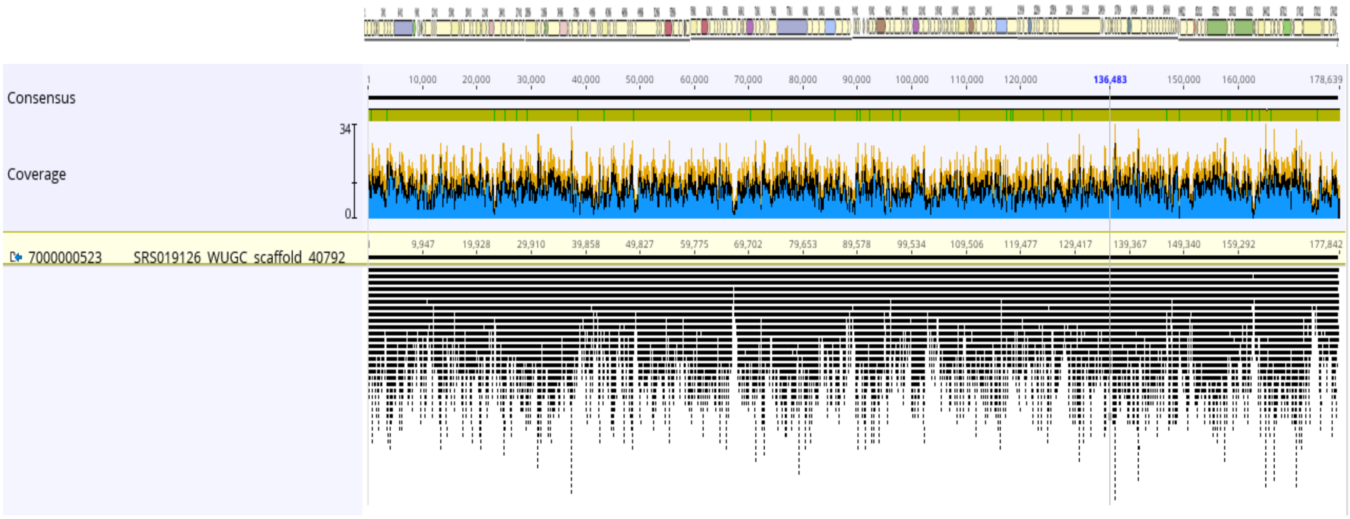
viralgroup 4000 (mean coverage 27.1, SD 12.5)



viralgroup 3591 (mean coverage 23.5, SD 7.3)



viralgroup 2981 (mean coverage 13.1, SD 4.7)



viralgroup 3836 (mean coverage 16.0, SD 4.7)

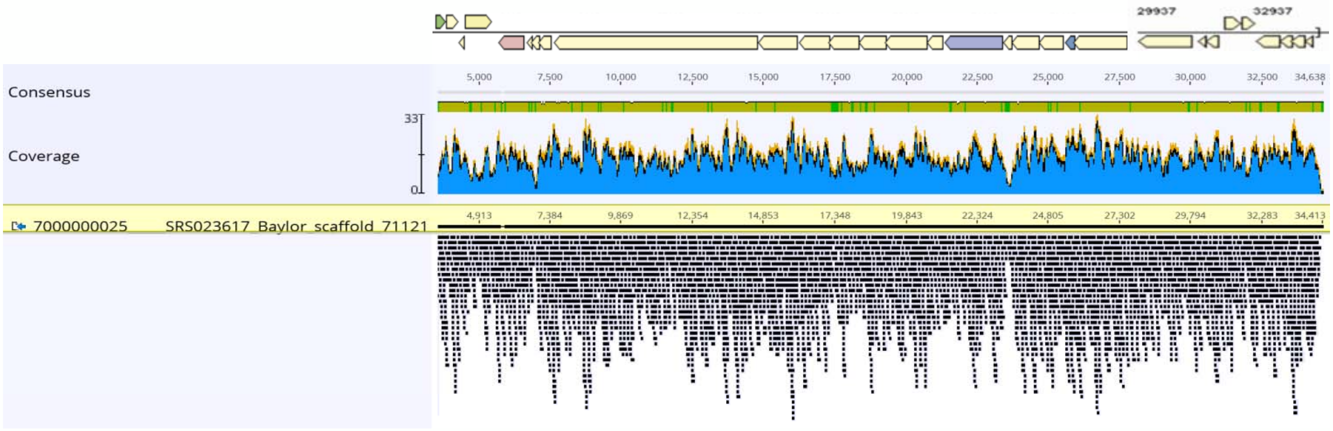
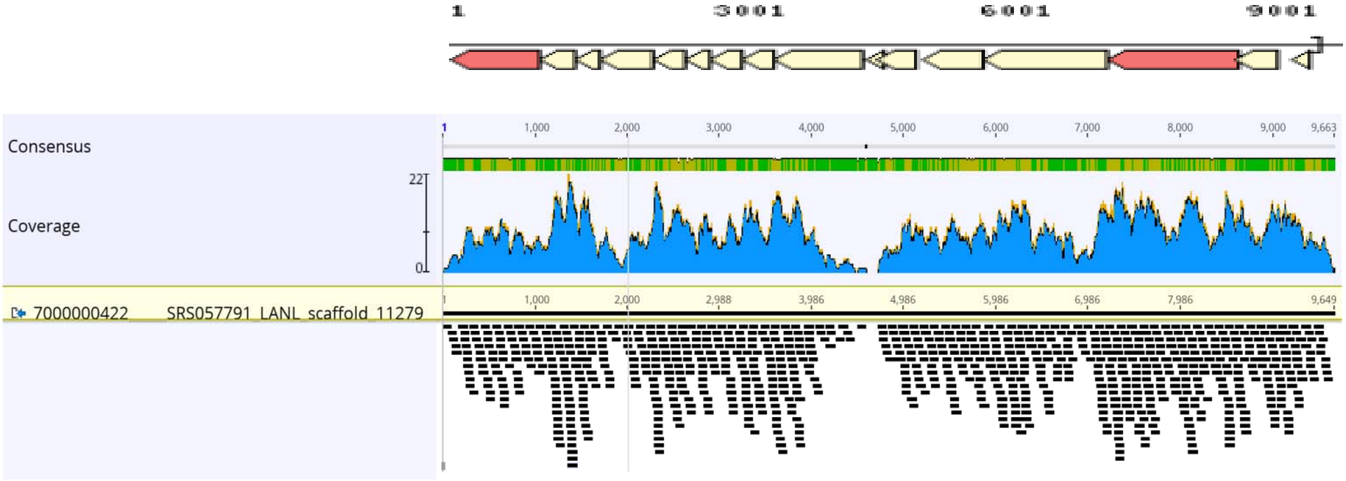
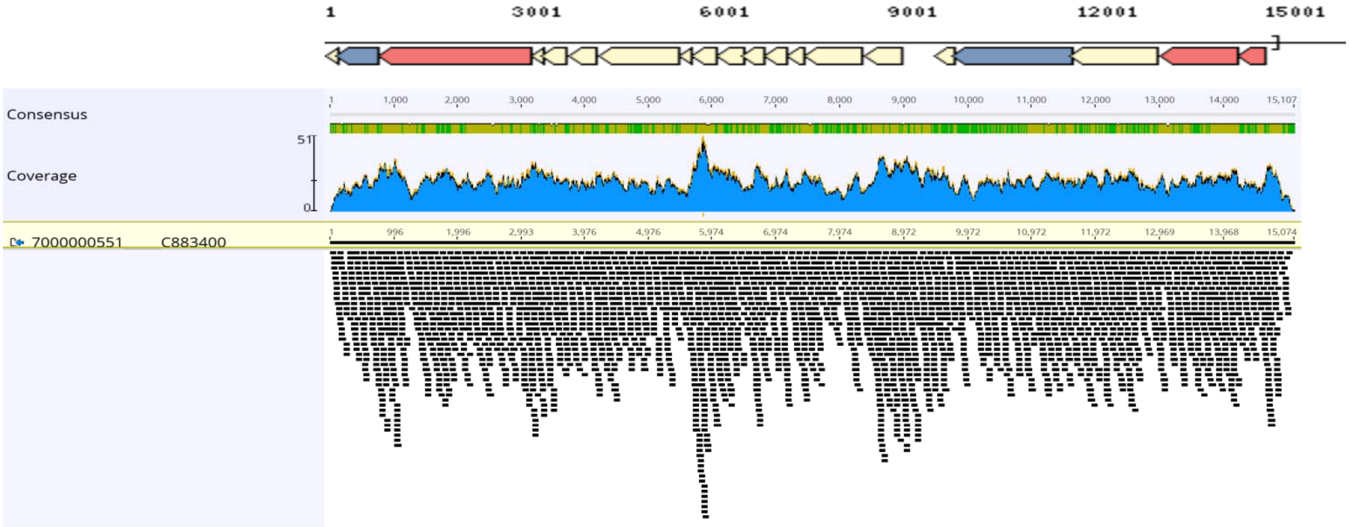


Figure S3

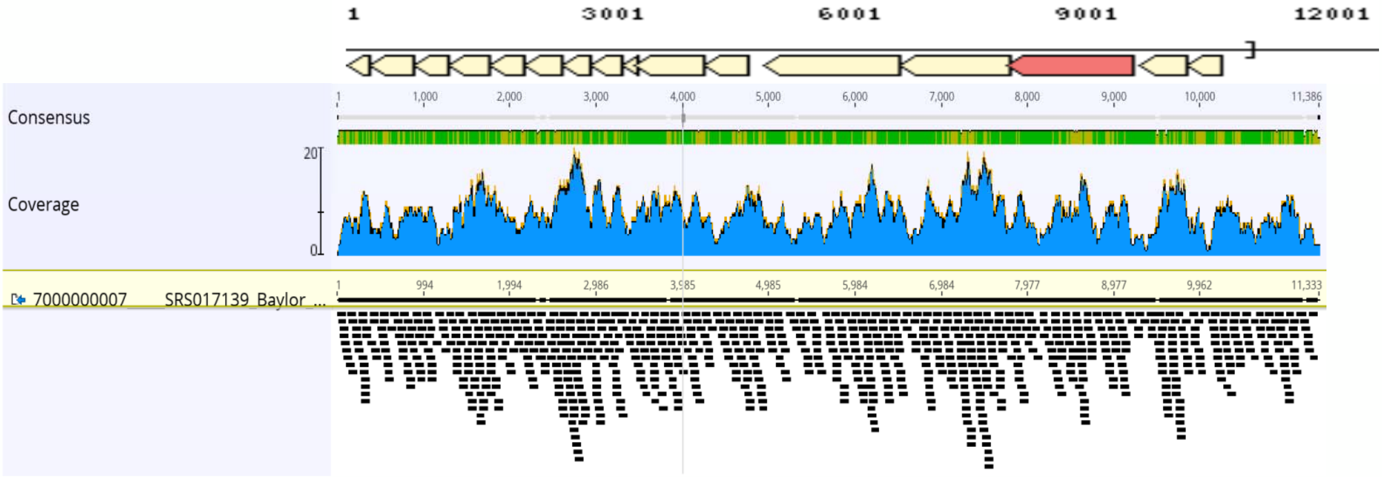
viralgroup 3050 (mean coverage 9.0, SD 4.2)



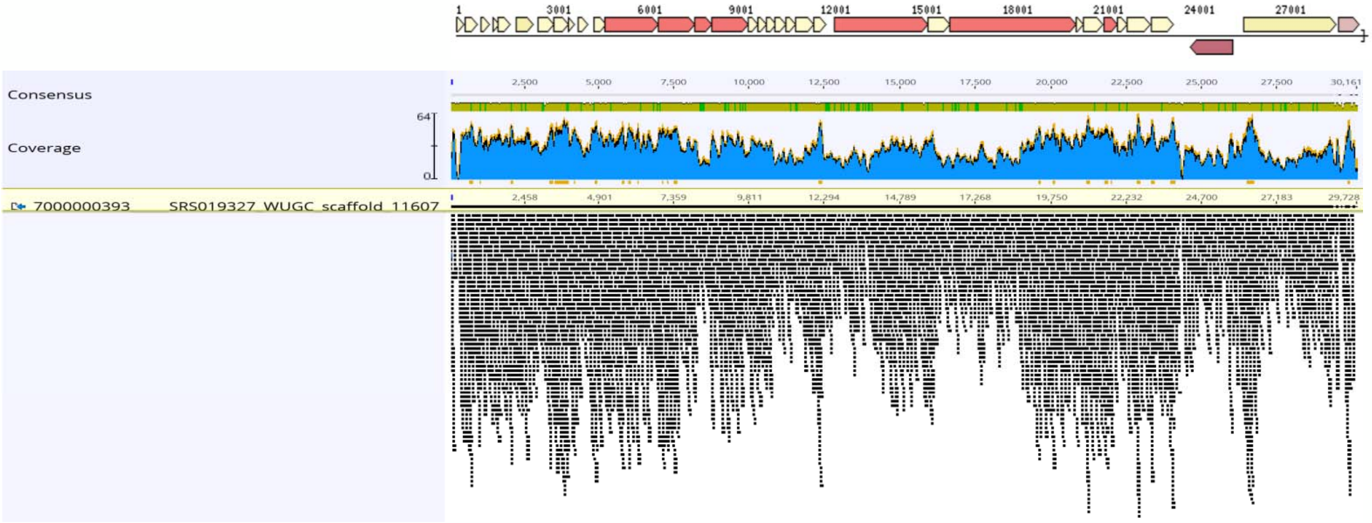
viralgroup 3445 (mean coverage 20.8, SD 5.7)



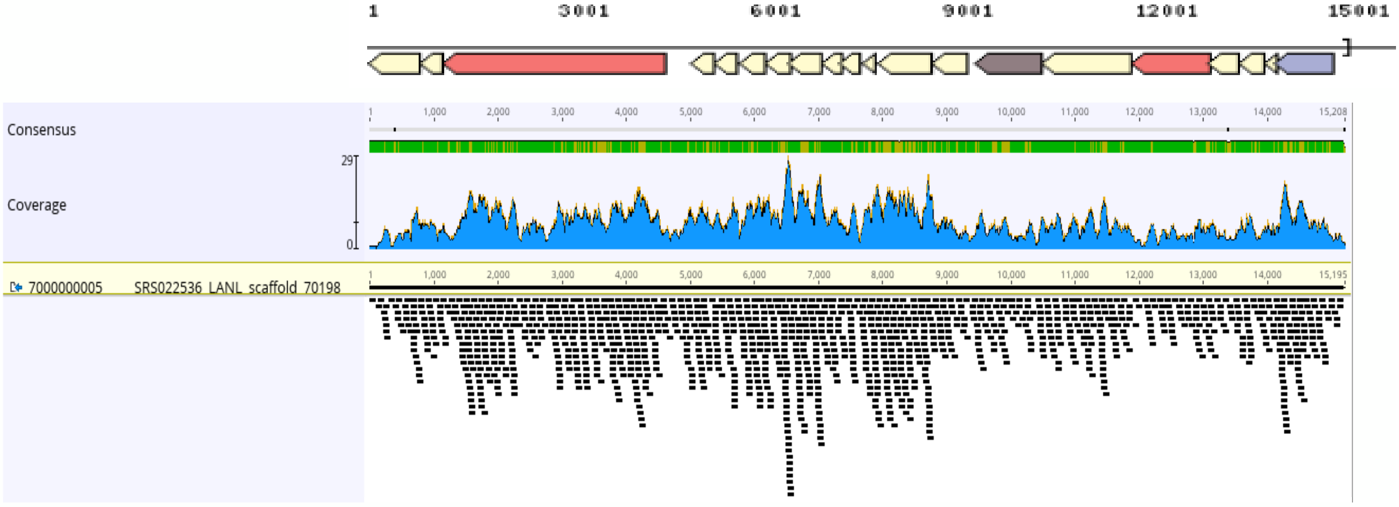
viralgroup 4484 (mean coverage 21.2, SD 8.2)



viralgroup 3358 (mean coverage 32.0, SD 10.9)



viralgroup 3915 (mean coverage 8.5, SD 4.3)



viralgroup 3200 (mean coverage 44.7, SD 33.7)

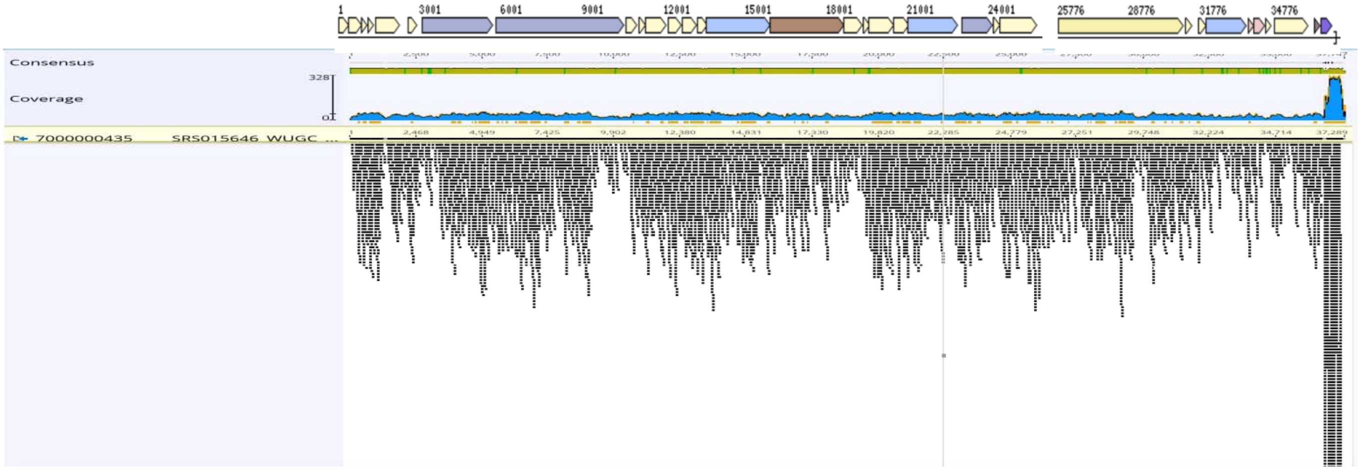
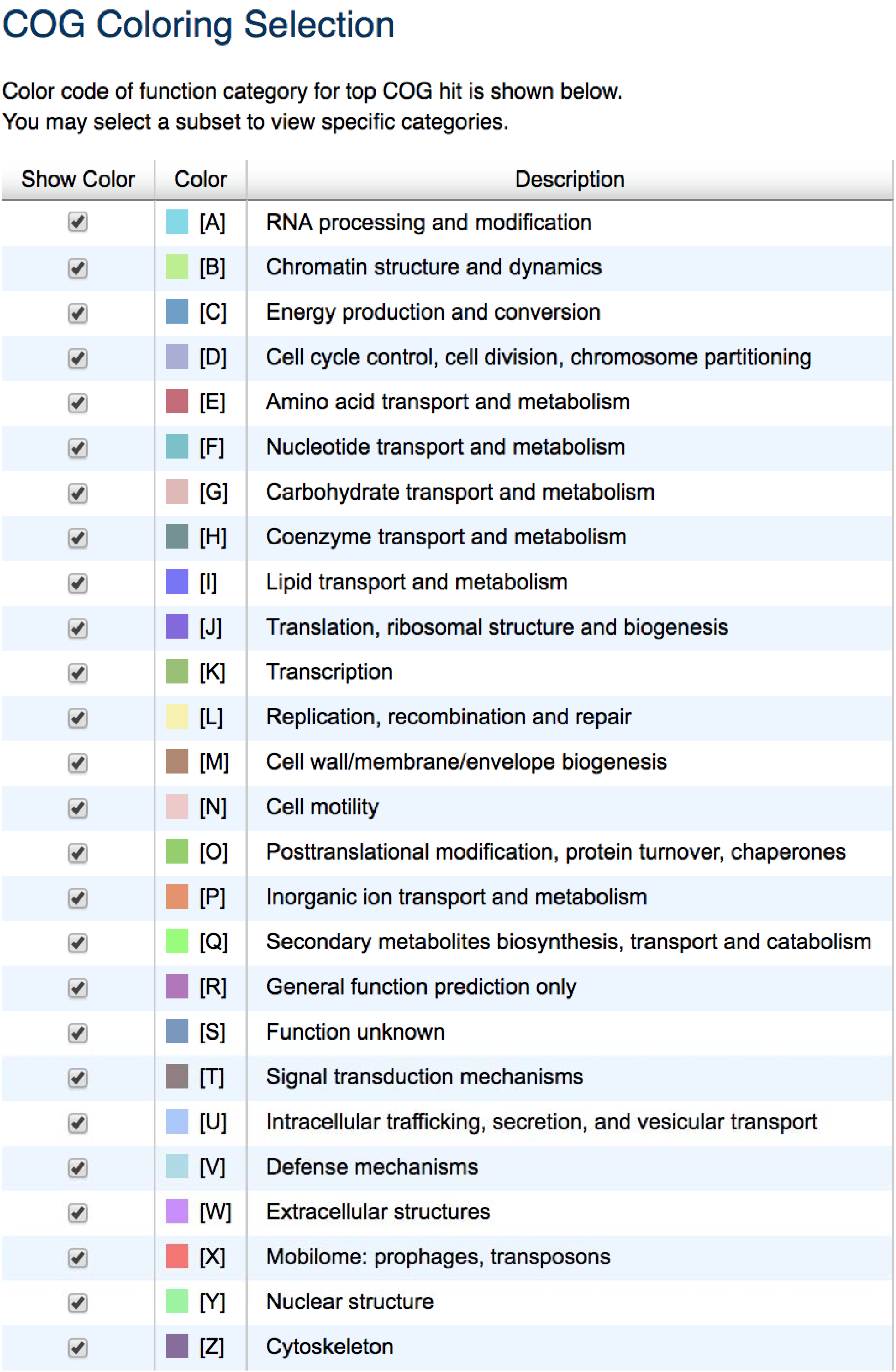


Figure S3



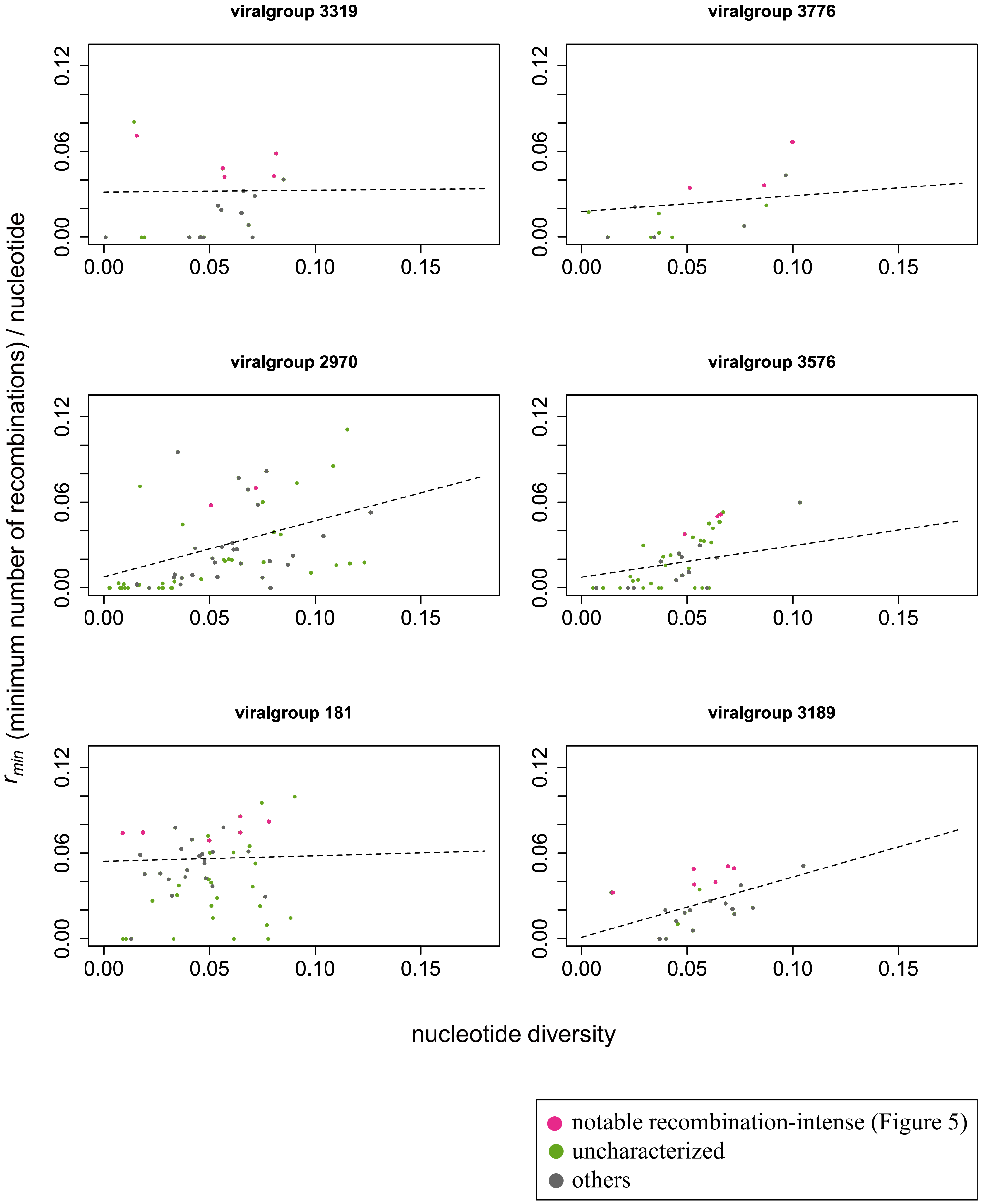
COG Coloring Selection

Color code of function category for top COG hit is shown below.
You may select a subset to view specific categories.

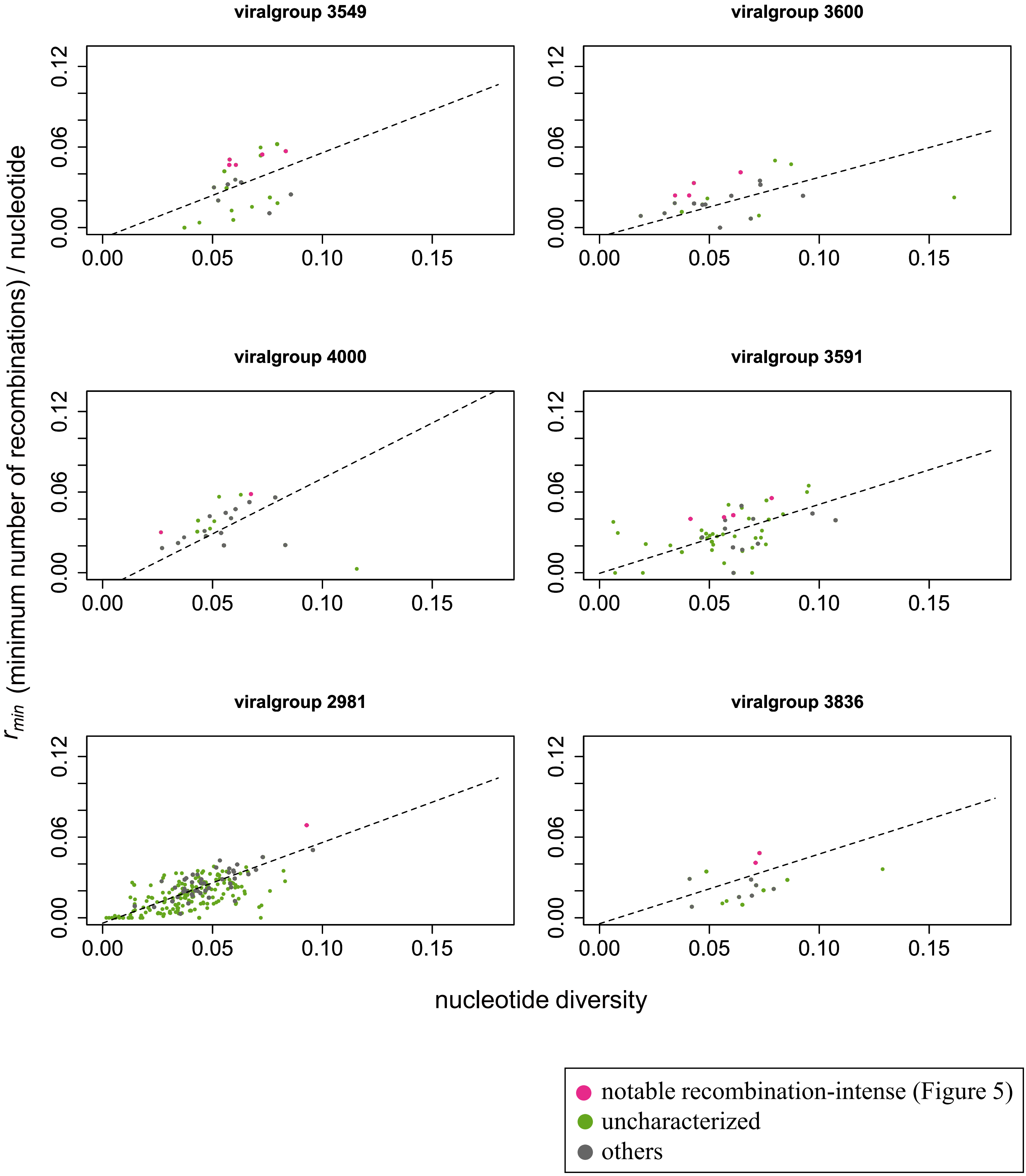
Show Color	Color	Description
☑	🟦 [A]	RNA processing and modification
☑	🟩 [B]	Chromatin structure and dynamics
☑	🟦 [C]	Energy production and conversion
☑	🟪 [D]	Cell cycle control, cell division, chromosome partitioning
☑	🟥 [E]	Amino acid transport and metabolism
☑	🟦 [F]	Nucleotide transport and metabolism
☑	🟥 [G]	Carbohydrate transport and metabolism
☑	⬛ [H]	Coenzyme transport and metabolism
☑	🟦 [I]	Lipid transport and metabolism
☑	🟪 [J]	Translation, ribosomal structure and biogenesis
☑	🟩 [K]	Transcription
☑	🟨 [L]	Replication, recombination and repair
☑	🟫 [M]	Cell wall/membrane/envelope biogenesis
☑	🟥 [N]	Cell motility
☑	🟩 [O]	Posttranslational modification, protein turnover, chaperones
☑	🟧 [P]	Inorganic ion transport and metabolism
☑	🟩 [Q]	Secondary metabolites biosynthesis, transport and catabolism
☑	🟪 [R]	General function prediction only
☑	🟦 [S]	Function unknown
☑	🟫 [T]	Signal transduction mechanisms
☑	🟦 [U]	Intracellular trafficking, secretion, and vesicular transport
☑	🟦 [V]	Defense mechanisms
☑	🟪 [W]	Extracellular structures
☑	🟥 [X]	Mobilome: prophages, transposons
☑	🟩 [Y]	Nuclear structure
☑	🟪 [Z]	Cytoskeleton

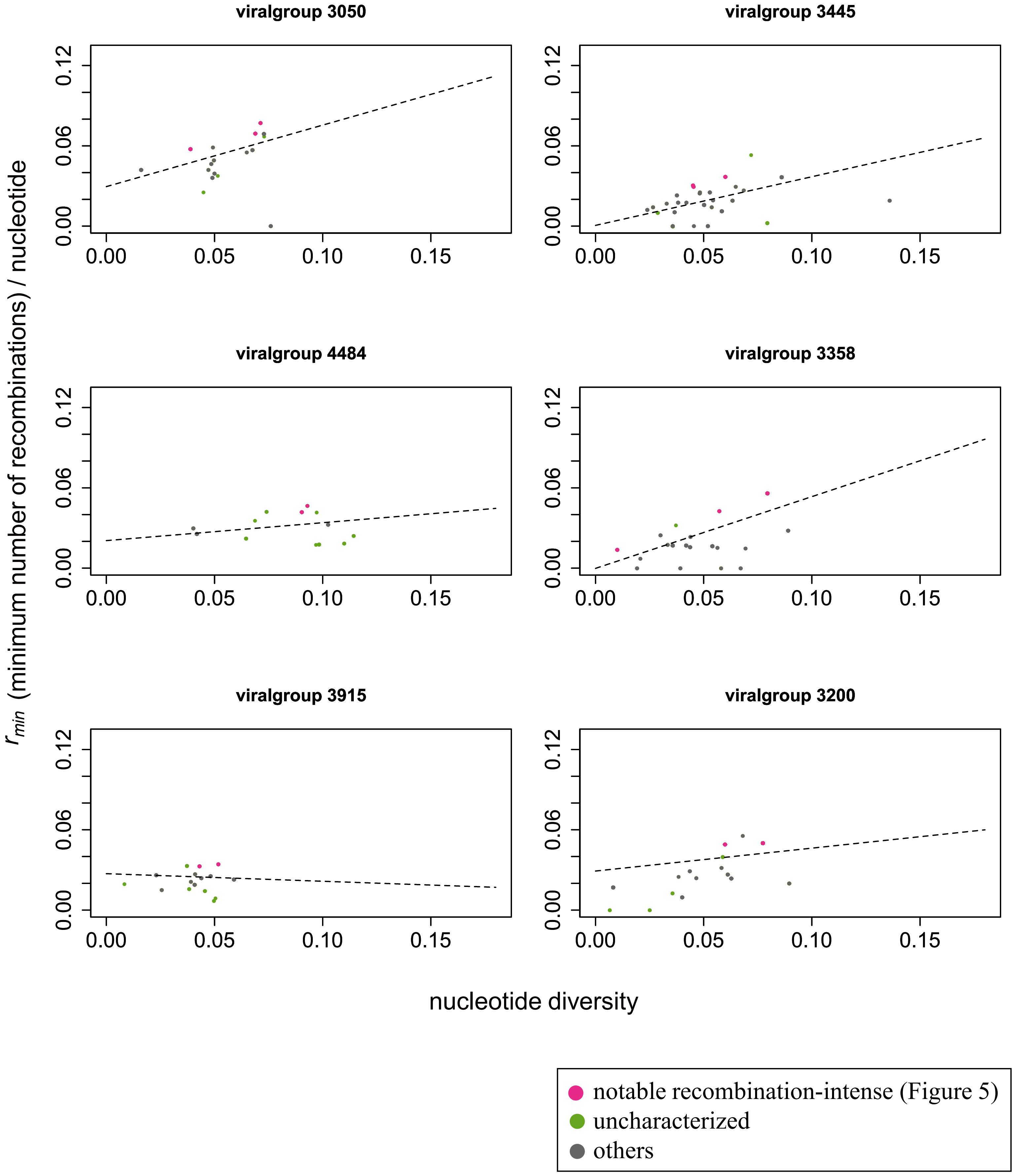
Tree scale: 0.1
① Species (ICTV)
② Species cluster (VICTOR)
③ Genus (ICTV)
3alikevirus
77likevirus
Ahjdlikevirus
Allolevivirus
Alphafusellovirus
Alphalipothrixvirus
Ampullavirus
Andromedalikevirus
Barnyardlikevirus
Bcep22likevirus
Bcep78likevirus
Bcepmulikevirus
Bdellomicrovirus
Betafusellovirus
Betalipothrixvirus
Bicaudavirus
Bignuzlikevirus
Bppunalikevirus
Bronlikevirus
C2likevirus
C5likevirus
Charlielikevirus
Che8likevirus
Che9clikevirus
Chilikevirus
Chlamydiamicrovirus
Cjwunalikevirus
Corndoglikevirus
Corticovirus
Cp220likevirus
Cp8unalikevirus
D3112likevirus
D3likevirus
Deltalipothrixvirus
Epsilon15likevirus
F116likevirus
Felixounalikevirus
Gammalipothrixvirus
Globulovirus
Halolikevirus
Hapunalikevirus
Hk578likevirus
Hpunalikevirus
Iebhlikevirus
Inovirus
Jerseylikevirus
L5likevirus
Lambdalikevirus
Levivirus
Luz24likevirus
Microvirus
Mulikevirus
N15likevirus
N4likevirus
Omegalikevirus
P22likevirus
P23likevirus
P2likevirus
Pbiunalikevirus
Pbunalikevirus
Pgonelikevirus
Phi29likevirus
Phic3unalikevirus
Phicbklikevirus
Phicd119likevirus
Phie125likevirus
Phieco32likevirus
Phietalikevirus
Phifllikevirus
Phijlunalikevirus
Phikmvlikevirus
Phikzlikevirus
Plasmavirus
Plectrovirus
Punalikevirus
Reylikevirus
Rudivirus
Sap6likevirus
Schizot4likevirus
Sfi1unalikevirus
Sfi21dtunalikevirus
Skunalikevirus
Sp6likevirus
Spbetalikevirus
Spiromicrovirus
Spounalikevirus
T4likevirus
T5likevirus
T7likevirus
Tectivirus
Tm4likevirus
Tp2unalikevirus
Tunalikevirus
Twortlikevirus
Unassigned
Viunalikevirus
Wbetalikevirus
Xp10likevirus
Yualikevirus
NA
④ Genus cluster (VICTOR)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
NA
⑤ Subfamily (ICTV)
Autographivirinae
Eucampyvirinae
Gokushovirinae
Peduovirinae
Picovirinae
Spounavirinae
Tevenvirinae
NA
⑥ Subfamily cluster (VICTOR)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
NA
⑦ Family (ICTV)
Ampullaviridae
Bicaudaviridae
Corticoviridae
Fuselloviridae
Globuloviridae
Inoviridae
Leviviridae
Lipothrixviridae
Microviridae
Myoviridae
Plasmaviridae
Podoviridae
Rudiviridae
Siphoviridae
Tectiviridae
NA
⑧ Family cluster (VICTOR)
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
NA
Species
①
Species
②
Genus
③④
Subfamily
⑤⑥
Family
⑦⑧
Bordetella phage BPP-1 (AY029185) 291
Endosymbiont phage APSE-1 (AF157835) 462
Vibrio phage CTX (GU942563) 390
Salmonella phage epsilon15 (AY150271) 298
Vibrio phage CTX (KJ540276) 390
Vibrio phage CTX (KJ540274) 390
Vibrio phage CTX (KJ540277) 390
Vibrio phage CTX (KJ540278) 390
Vibrio phage CTX (KJ540263) 390
Vibrio phage CTX (KJ540260) 390
Vibrio phage CTX (KJ540267) 390
Vibrio phage CTX (KJ540268) 390
Vibrio phage CTX (KJ540262) 390
Vibrio phage CTX (KJ540257) 390
Vibrio phage CTX (KJ540258) 390
Vibrio phage CTX (KJ540259) 390
Vibrio phage CTX (HQ224500) 390
Vibrio phage CTX (KF664579) 390
Vibrio phage CTX (KJ540266) 390
Vibrio phage CTX (KJ540270) 390
Vibrio phage CTX (KF664581) 390
Vibrio phage CTX (KM352500) 390
Vibrio phage CTX (KJ540261) 390
Vibrio phage CTX (KJ540264) 390
Vibrio phage CTX (KJ540271) 390
Vibrio phage CTX (KJ540272) 390
Vibrio phage CTX (KJ619459) 388
Enterobacteria phage BZ13 (NC001426) 467
Enterobacteria phage BZ13 (FJ483839) 467
Enterobacteria phage BZ13 (FJ483837) 467
Enterobacteria phage BZ13 (FJ483838) 465
Enterobacteria phage BZ13 (AF227250) 466
Enterobacteria phage St-1 (GQ149088) 44
Enterobacteria phage phiK (X60323) 44
85
Enterobacteria phage alpha3 (DQ085810) 39
52
Enterobacteria phage alpha3 (NC001330) 39
Enterobacteria phage G4 (V00657) 5
Enterobacteria phage G4 (AF454431) 5
Enterobacteria phage G4 (GQ153921) 5
Enterobacteria phage G4 (JF719730) 5
Enterobacteria phage G4 (GQ153922) 5
Enterobacteria phage G4 (JF719731) 5

Figure S5



viralgroup 3319
viralgroup 3776
viralgroup 2970
viralgroup 3576
viralgroup 181
viralgroup 3189
0.00
0.06
0.12
0.00
0.05
0.10
0.15
r_{min} (minimum number of recombinations) / nucleotide
nucleotide diversity
● notable recombination-intense (Figure 5)
● uncharacterized
● others


viralgroup 3549
viralgroup 3600
viralgroup 4000
viralgroup 3591
viralgroup 2981
viralgroup 3836
0.00
0.06
0.12
0.00
0.05
0.10
0.15
r_{min} (minimum number of recombinations) / nucleotide
nucleotide diversity
● notable recombination-intense (Figure 5)
● uncharacterized
● others


viralgroup 3050
viralgroup 3445
viralgroup 4484
viralgroup 3358
viralgroup 3915
viralgroup 3200
0.00
0.06
0.12
0.00
0.05
0.10
0.15
r_{min} (minimum number of recombinations) / nucleotide
nucleotide diversity
● notable recombination-intense (Figure 5)
● uncharacterized
● others