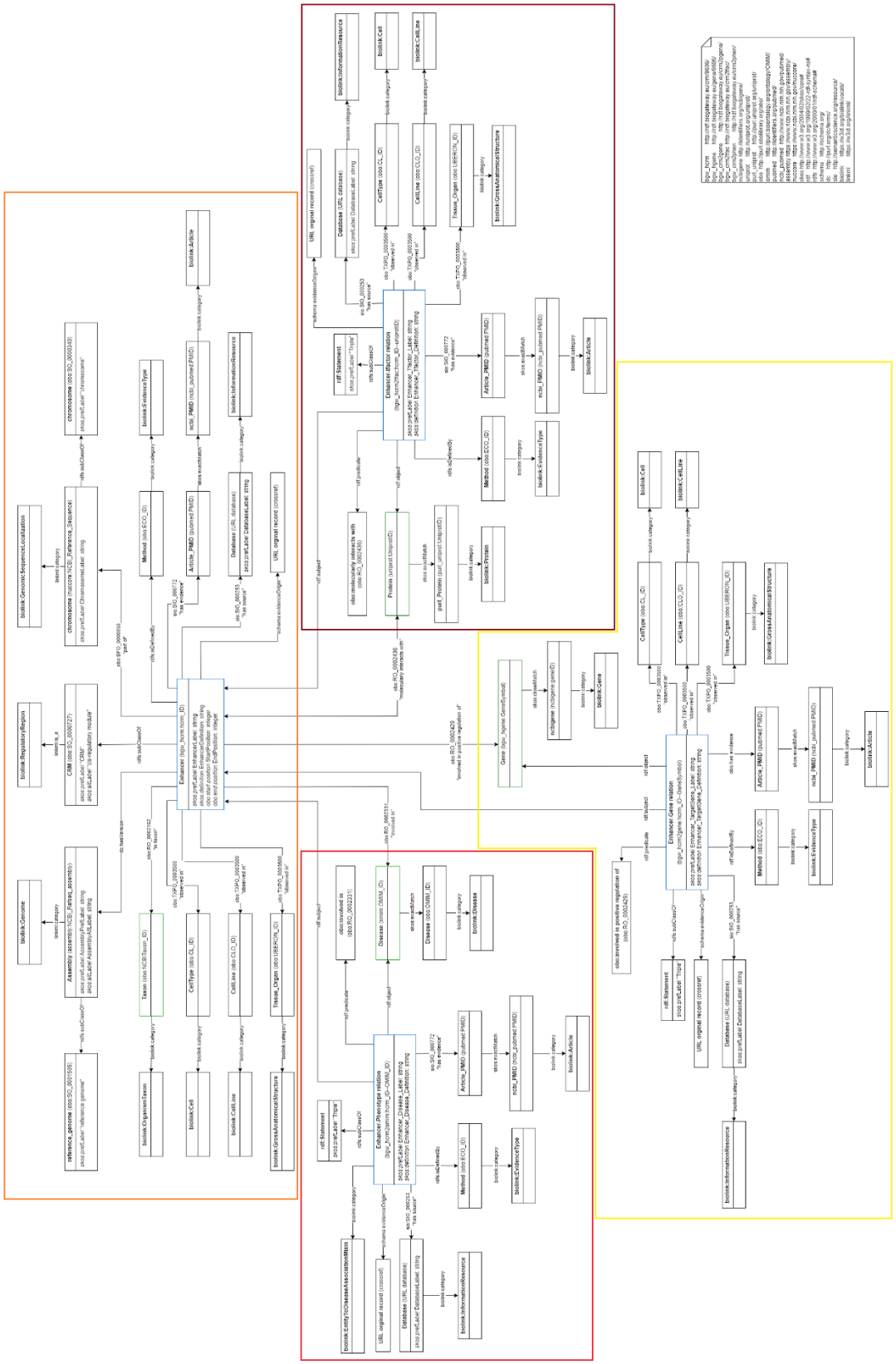


Figure 1. Schema used to model the information from different biological databases about enhancers and their relations with other entities. The colored boxes specify the different subgraphs or subdomains: enhancer sequences (crm graph - orange on the top), and their relations with other biological entities of interest, such as target genes (crm2gene graph - yellow on the bottom), transcription factors (crm2tfac graph - brown on the right), and phenotypes (crm2phen graph - red on the left). The blue classes constitute the central entities of each graph, while the green classes are biological classes of interest that were not modeled in detail because they are already present in the BioGateway KG, schema that is interoperable with this one.

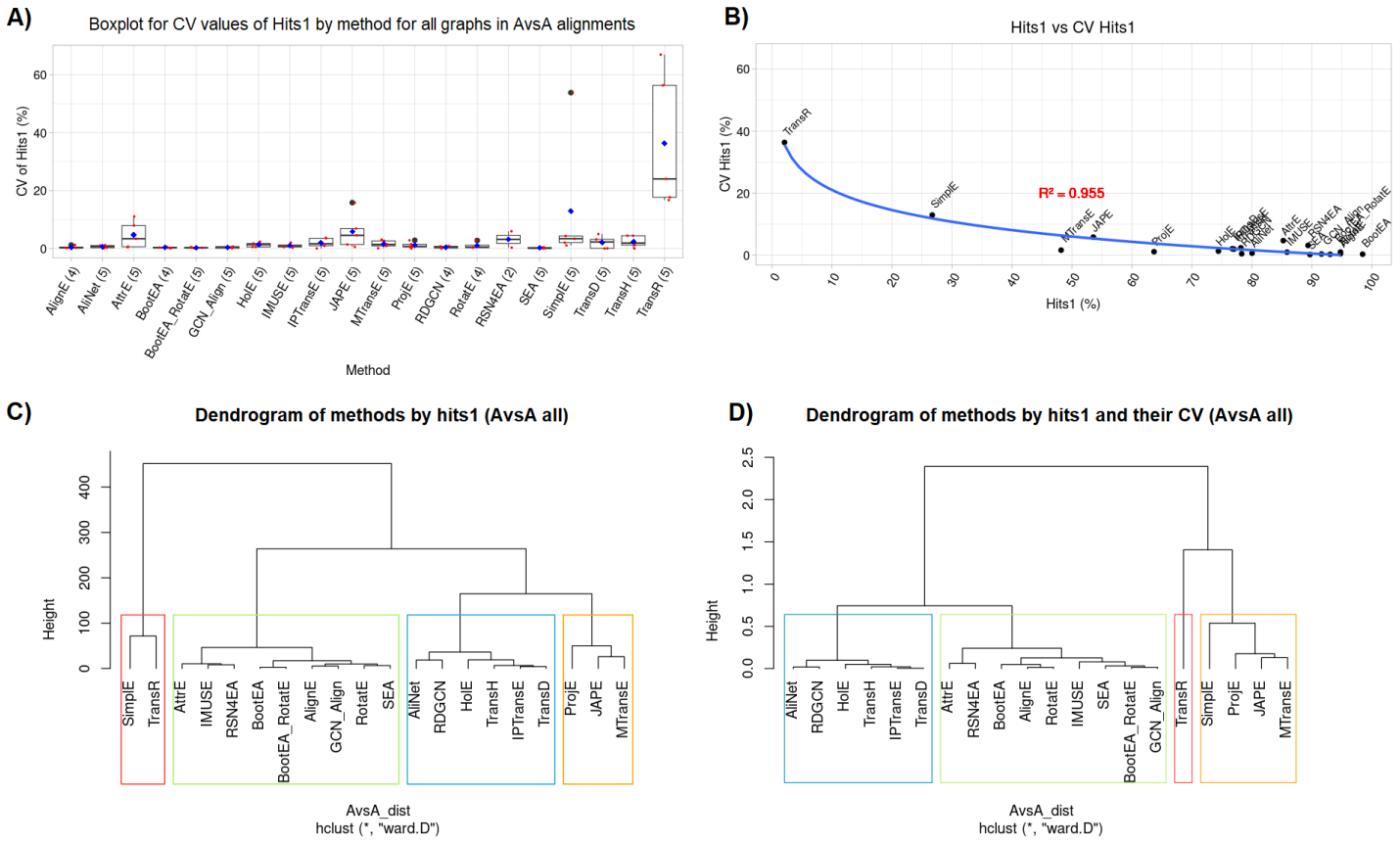


Figure 2. Results associated to hits@1 values in AvsA alignments with all data, and considering 5 datasets (ENdb, EnDisease, DiseaseEnhancer, VISTA and RefSeq). A) Boxplot for the coefficients of variation (CV) of hits@1 values between replicates for each method. B) Observed log trend between the hits@1 values and their CV, derived by running replicates. C and D) Comparison of methods classified from dendrograms using hits@1 values (C) and using hits@1 values and the CV of hits@1 values between replicates (D). The clusters remain mainly without changes, with two clusters (red and orange) associated with methods with low hits@1 values, one cluster (green) with good results, that could be subdivided in two subclusters, and one cluster with intermediate results (blue).

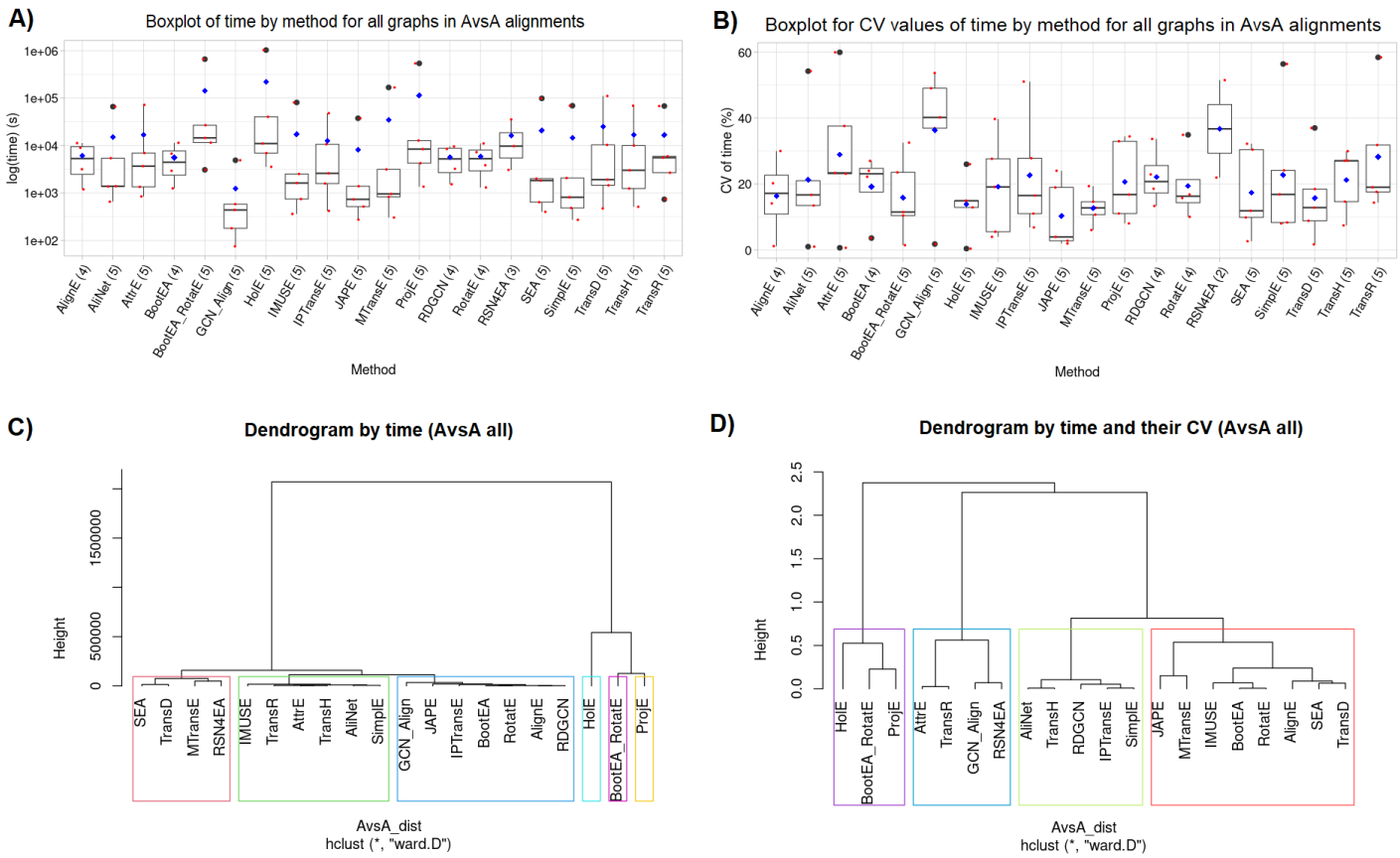


Figure 3. Results associated to time values in AvsA alignments with all data and considering 5 datasets (ENdb, EnDisease, DiseaseEnhancer, VISTA and RefSeq). A) Boxplot by method of time values. B) Boxplot of CV values of time. C) Dendrogram of methods using time values. Two main branches stand out. The left branch is associated with low run times and the right branch with high run times. These branches can in turn be subdivided into sub-clusters with more specific time profiles. D) Dendrogram of methods using time values and their CV. The inclusion of CV values of time does not improve the generation of clusters with homogeneous methods.

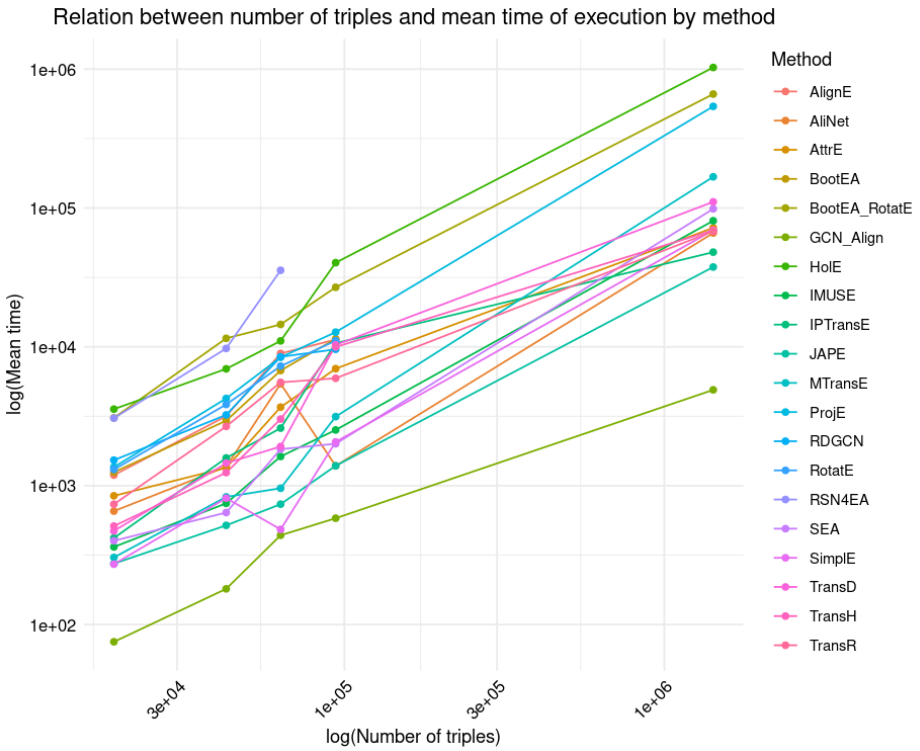


Figure 4. Representation of the number of triples against execution time for each alignment method used and each database. AvsA alignments with all data.

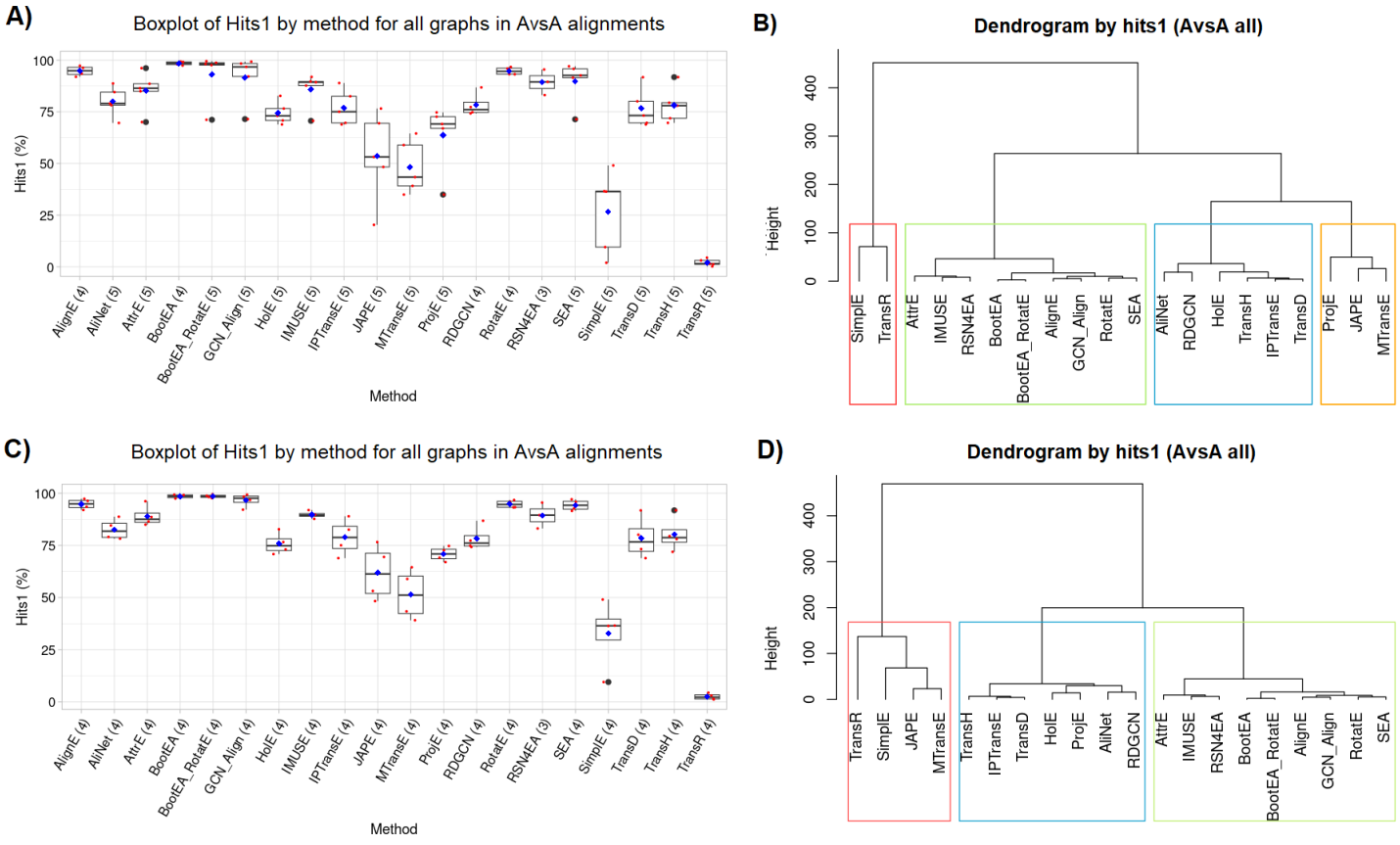


Figure 5. The hits@1 of the AvsA alignments with all data using five datasets (ENdb, EnDisease, DiseaseEnhancer, VISTA and RefSeq) (A and B) are compared to those obtained when RefSeq is removed due to non-alignments in some methods (C and D).

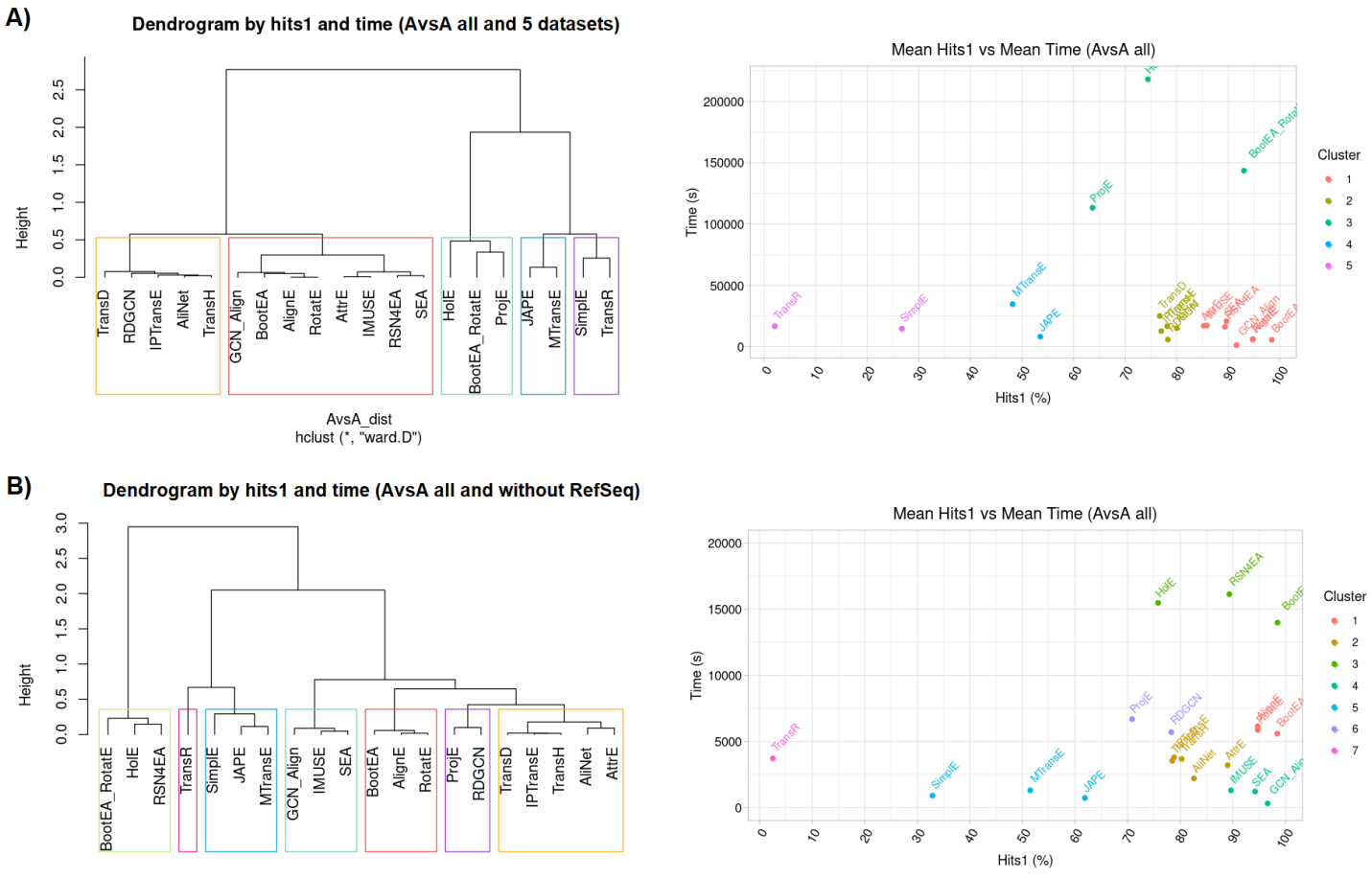
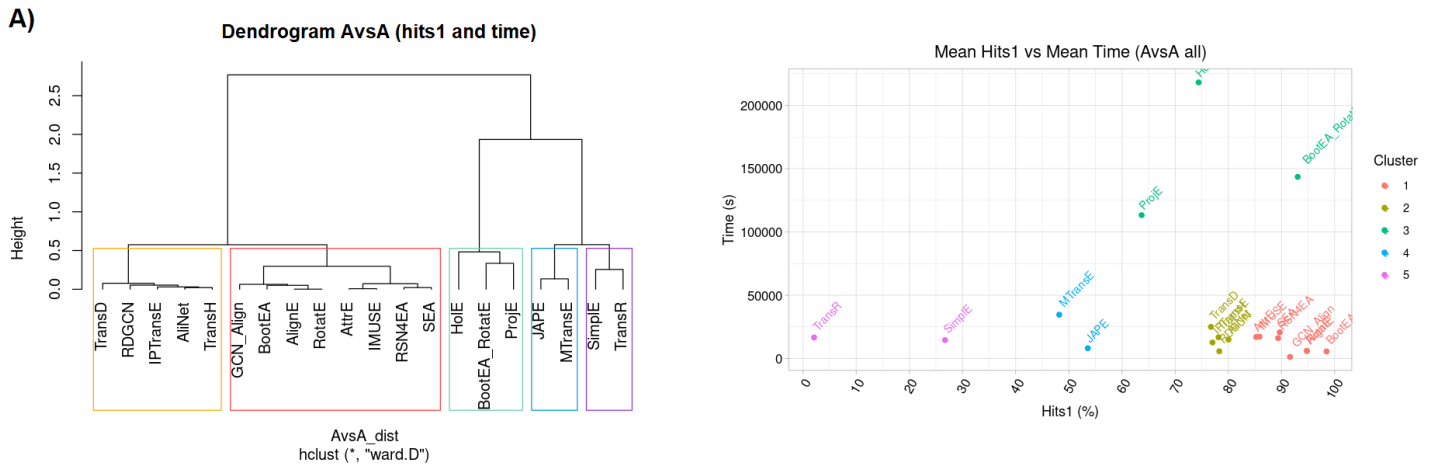


Figure 6. Dendrograms and representation of the clusters generated by grouping the methods by hits@1 and time values obtained in the AvsA alignments with all data. A) Results obtained by using the alignments of the 5 databases. B) Results when we omit the alignment corresponding to RefSeq, because different methods were not able to complete the alignment corresponding to this dataset.



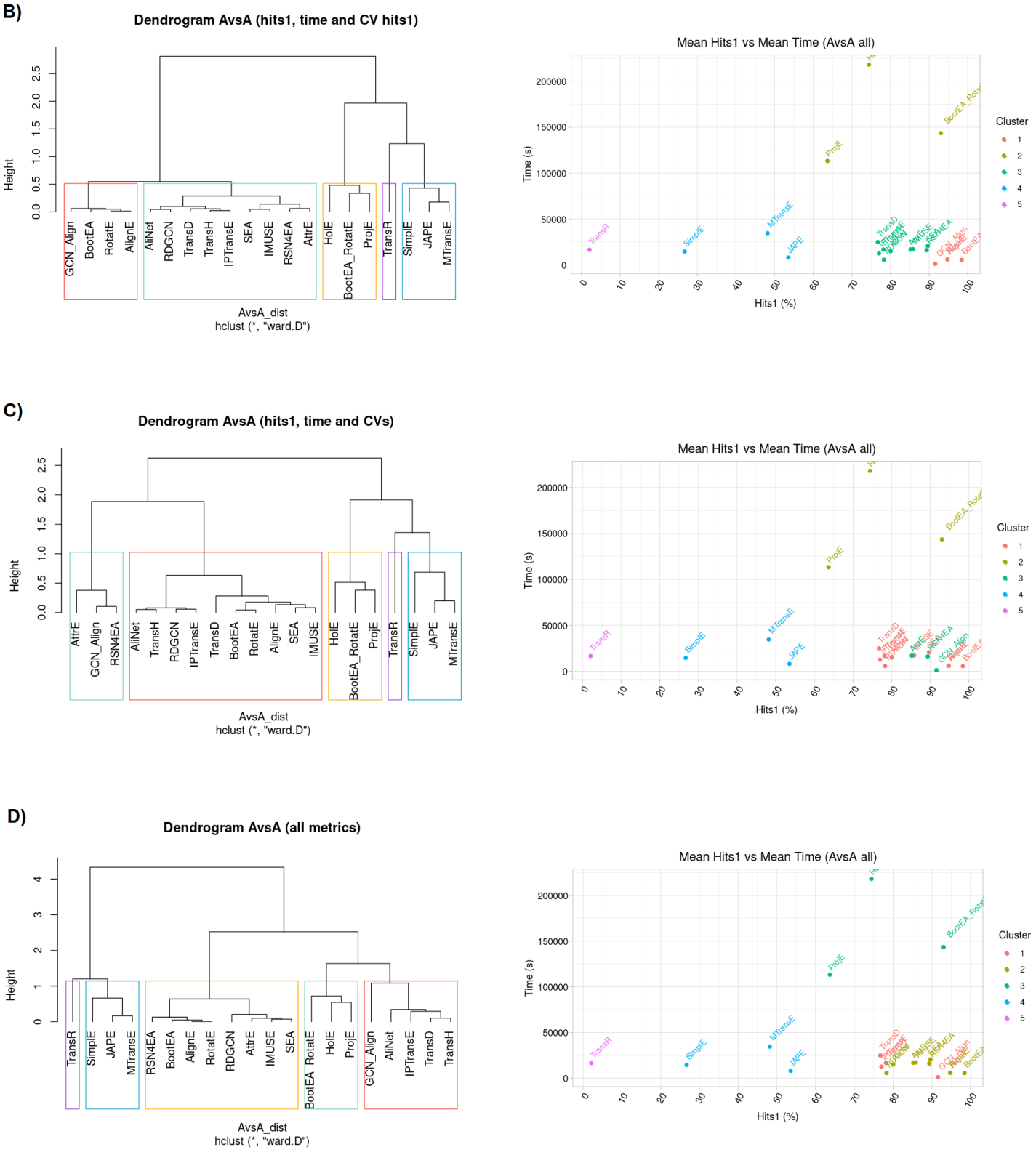
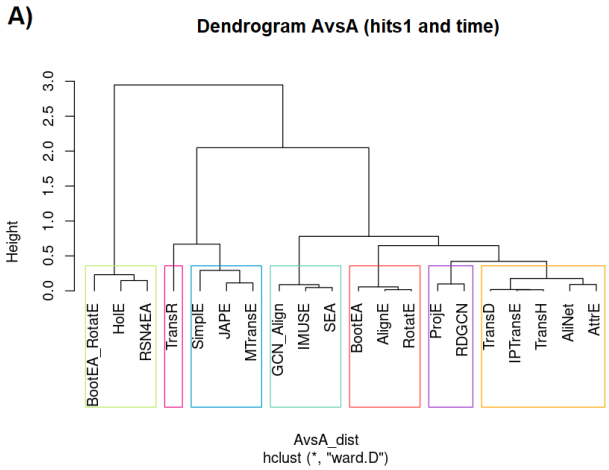
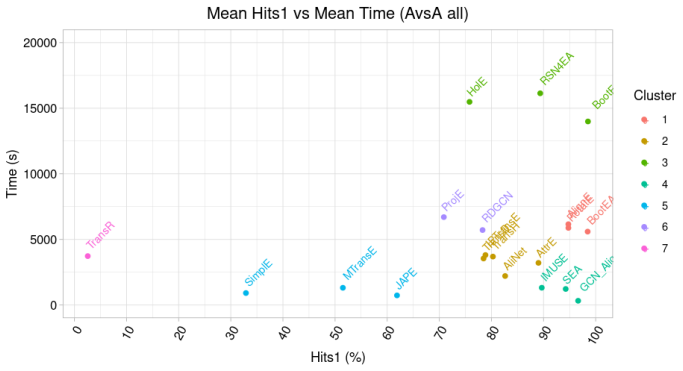


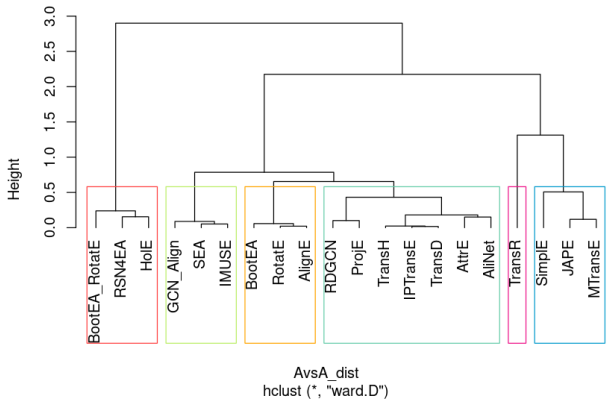
Figure 7. Effects of including coefficient of variance values of hits@1 and time, and other combination of scaled alignment metrics (hits@1, hits@5, hits@10, mr, mrr, time) in the clustering of methods in AvsA alignments with all data. A) Dendrogram and plot using hits@1 and time values. B) Dendrogram and plot after adding CV values of hits@1. C) Dendrogram and plot after adding CV values of hits@1 and time. D) Dendrogram and plot considering all metrics returned in the alignment (hits@1, hits5, hits@10, mr, mrr and time).

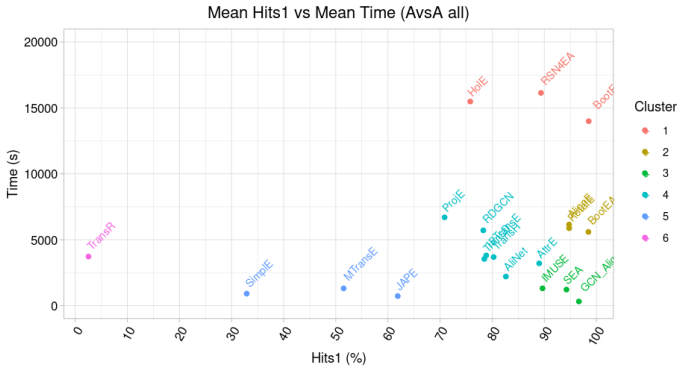
A)


Dendrogram AvsA (hits1 and time)
Height
BootEA_RotatE
HolE
RSN4EA
TransR
SimplE
JAPE
MTransE
GCN_Align
IMUSE
SEA
BootEA
AlignE
RotatE
ProjE
RDGCN
TransD
IPTransE
TransH
AliNet
AttrE
AvsA_dist
hclust (*, "ward.D")

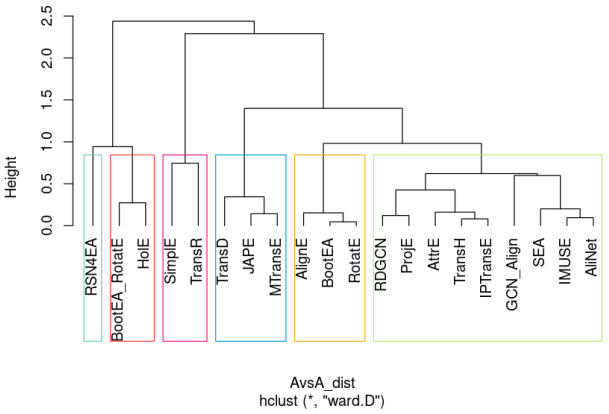

Mean Hits1 vs Mean Time (AvsA all)
Time (s)
Hits1 (%)
Cluster

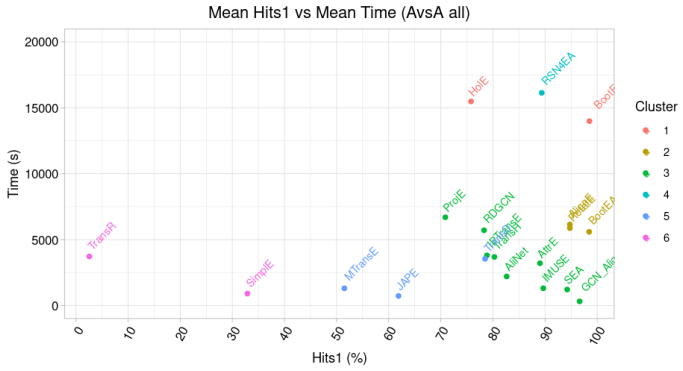
B)


Dendrogram AvsA (hits1, time and CV hits1)
Height
BootEA_RotatE
RSN4EA
HolE
GCN_Align
SEA
IMUSE
BootEA
RotatE
AlignE
RDGCN
ProjE
TransH
IPTransE
TransD
AttrE
AliNet
TransR
SimplE
JAPE
MTransE
AvsA_dist
hclust (*, "ward.D")

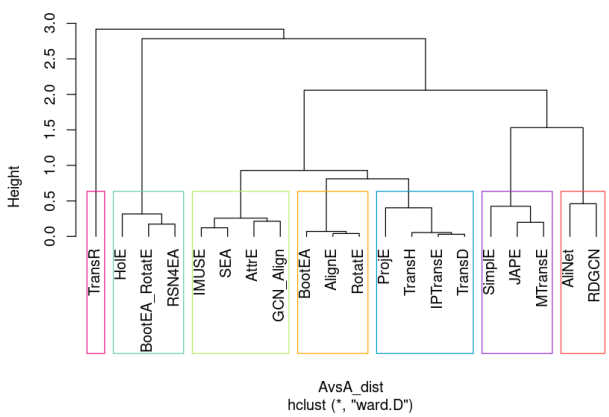

Mean Hits1 vs Mean Time (AvsA all)
Time (s)
Hits1 (%)
Cluster

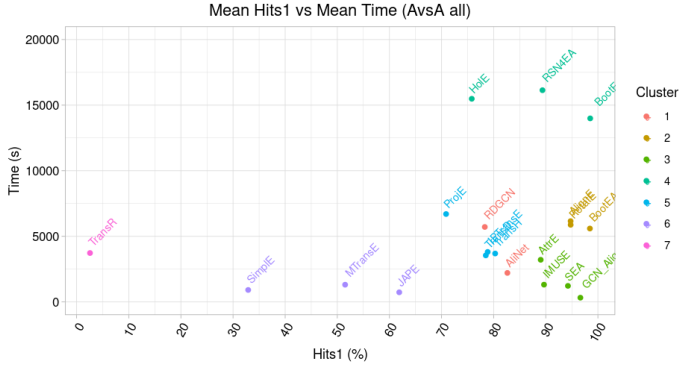
C)


Dendrogram AvsA (hits1, time and CVs)
Height
RSN4EA
BootEA_RotatE
HolE
SimplE
TransR
TransD
JAPE
MTransE
AlignE
BootEA
RotatE
RDGCN
ProjE
AttrE
TransH
IPTransE
GCN_Align
SEA
IMUSE
AliNet
AvsA_dist
hclust (*, "ward.D")


Mean Hits1 vs Mean Time (AvsA all)
Time (s)
Hits1 (%)
Cluster

D)


Dendrogram AvsA (all metrics)
Height
TransR
HolE
BootEA_RotatE
RSN4EA
IMUSE
SEA
AttrE
GCN_Align
BootEA
AlignE
RotatE
ProjE
TransH
IPTransE
TransD
SimplE
JAPE
MTransE
AliNet
RDGCN
AvsA_dist
hclust (*, "ward.D")


Mean Hits1 vs Mean Time (AvsA all)
Time (s)
Hits1 (%)
Cluster

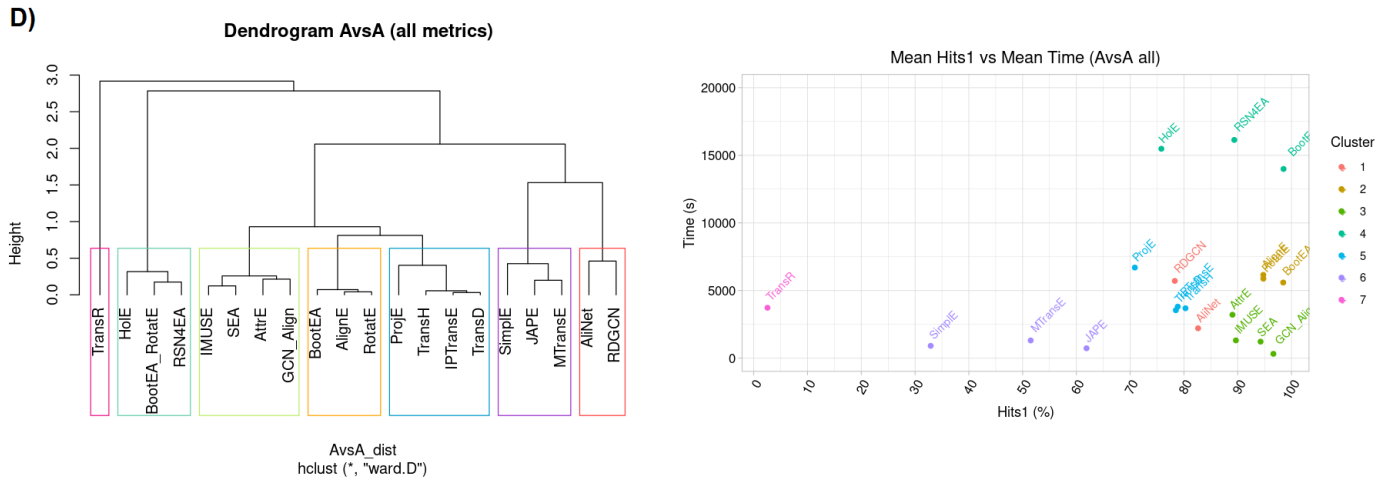
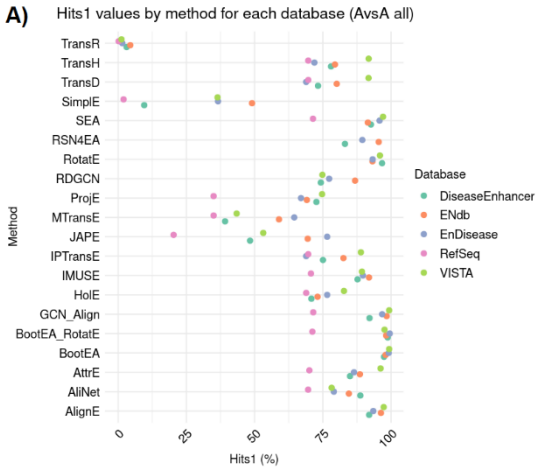


Figure 8. Effects of including coefficient of variance values of hits@1 and time, and other combination of scaled alignment metrics (hits@1, hits@5, hits@10, mr, mrr, time) in the clustering of methods in AvsA alignments with all data. The results of RefSeq were omitted from this report because different methods were not able to align this dataset. A) Dendrogram and plot using hits@1 and time values. B) Dendrogram and plot after adding CV values of hits@1. C) Dendrogram and plot after adding CV values of hits@1 and time. D) Dendrogram and plot considering all metrics returned in the alignment (hits@1, hits5, hits@10, mr, mrr and time).



B)

	database	first	second	third	fourth	fifth	weight
1	DiseaseEnhancer	1	1	1	5	0	4.4
2	ENdb	1	3	2	2	0	5.4
3	EnDisease	1	3	4	0	0	5.8
4	RefSeq	0	0	0	0	8	1.6
5	VISTA	5	1	1	1	0	6.8

C)

	database	first	second	third	fourth	fifth	weight
1	DiseaseEnhancer	0	1	1	3	0	2.6
2	ENdb	1	2	1	1	0	3.6
3	EnDisease	1	2	2	0	0	3.8
4	RefSeq	0	0	0	0	5	1
5	VISTA	3	0	1	1	0	4

Figure 9. A) Plot of hits@1 values by method and dataset in AvsA all type alignments. B) Ranking of the databases using the relative position of their hits@1 value when considering the 8 best-performing methods after clustering by hits@1 and time (GCN Align, SEA, IMUSE, BootEA, AlignE, RotatE, AttrE and BootEA-RotatE). C) Ranking of the databases using the relative position of their hits@1 value when considering the 5 best-performing methods after clustering by hits@1 and time (GCN Align, SEA, IMUSE, AttrE and BootEA-RotatE).

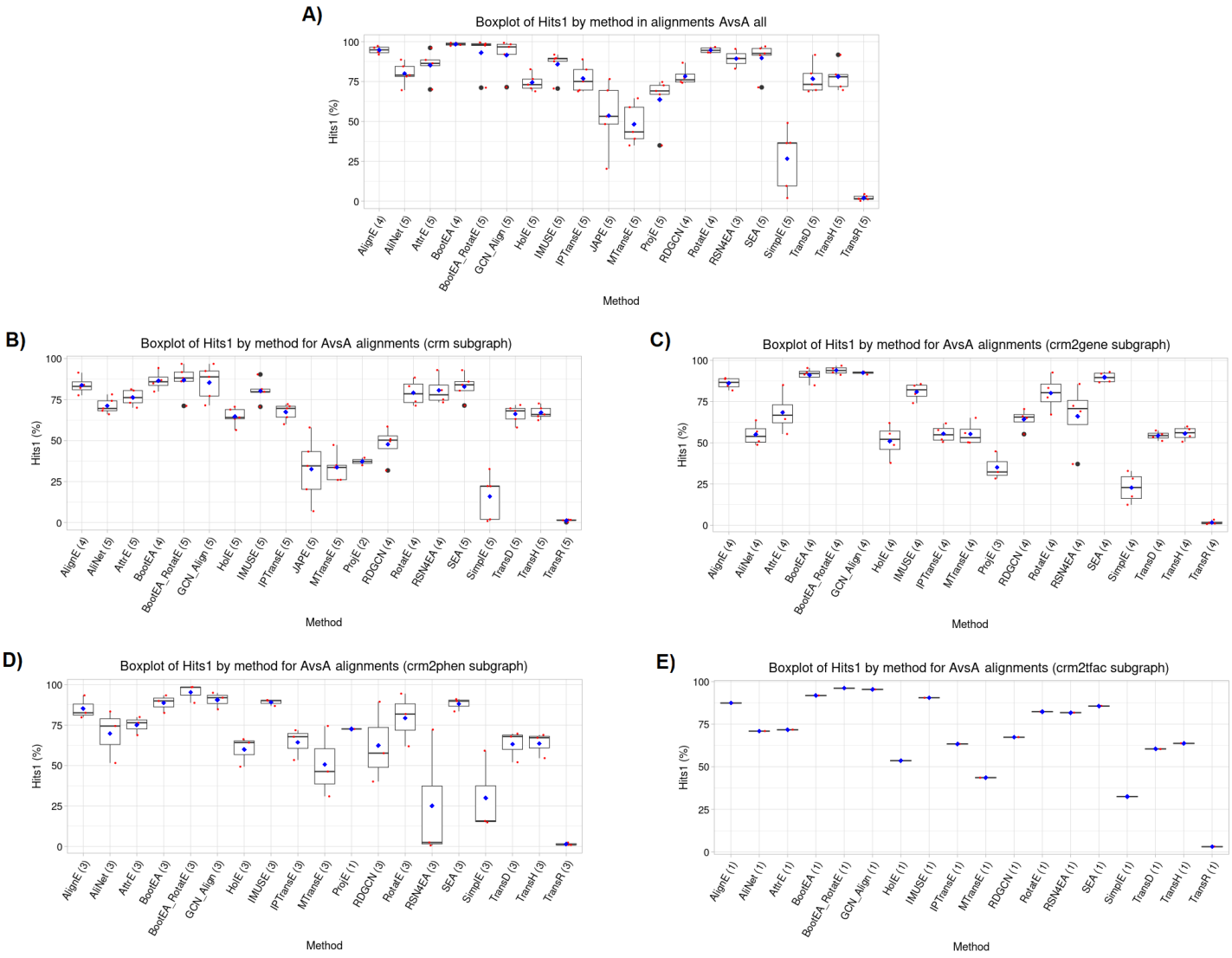


Figure 10. Boxplot of hits@1 values obtained in AvsA alignment using all the information from the source database (A), and different data subdomains: crm sequences (B), relations between crm and genes (C), relations between crm and phenotypes (D), and relations between crm and transcription factors (E).

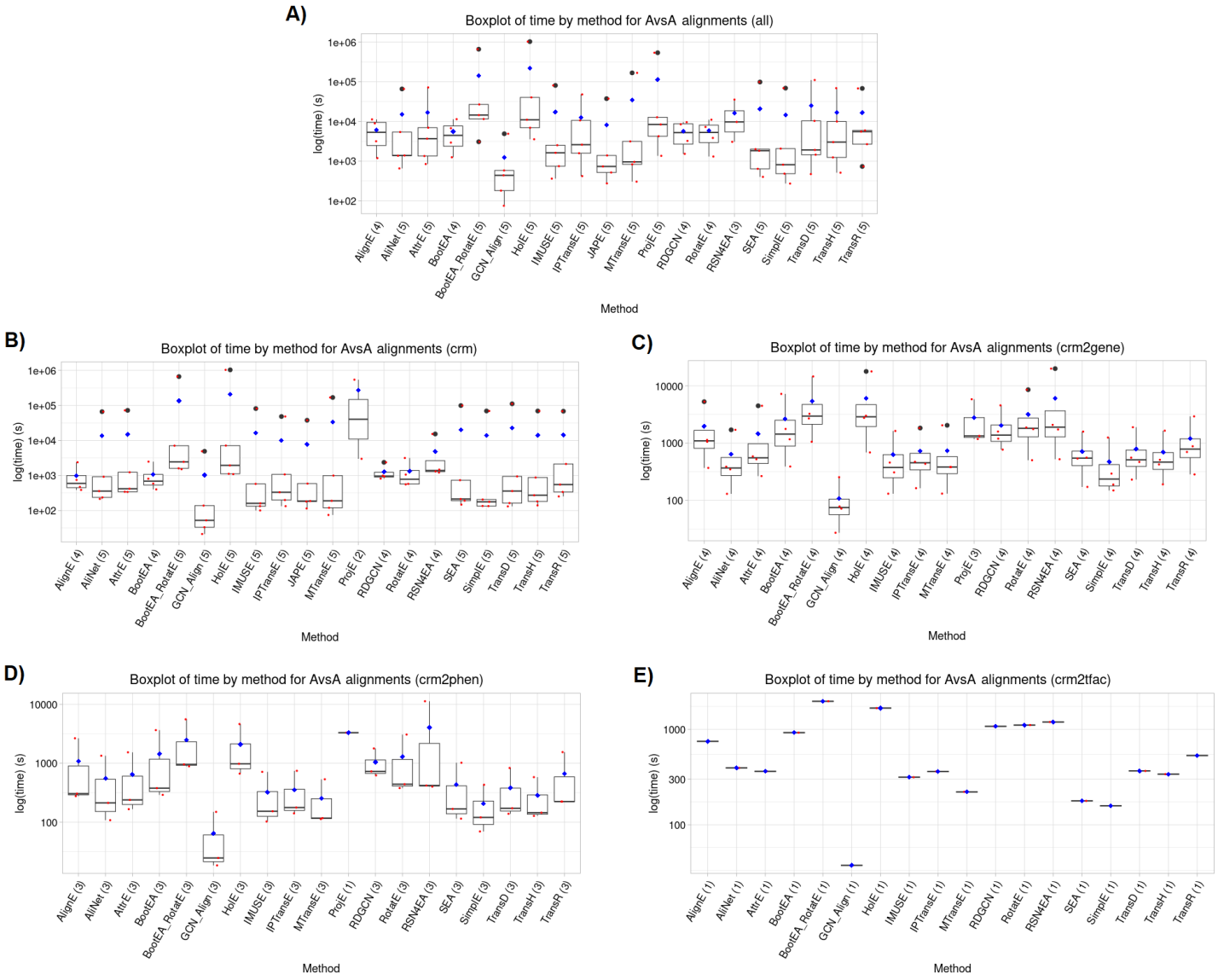


Figure 11. Boxplot of time values obtained in AvsA alignment using all the data domains from the source database (A), and different data subdomains: crm sequences (B), relations between crm and genes (C), relations between crm and phenotypes (D), and relations between crm and transcription factors (E).

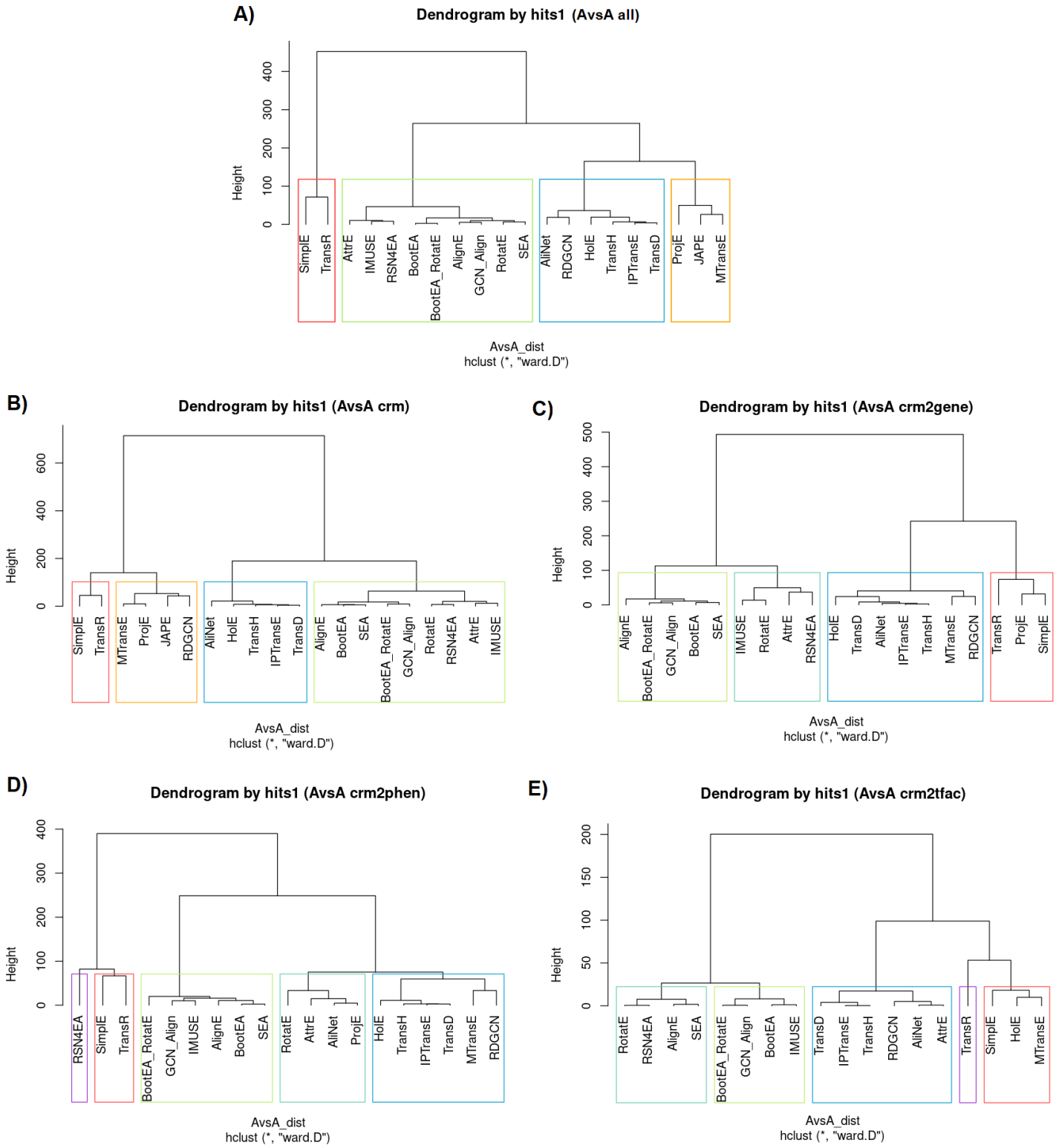


Figure 12. Dendrograms of methods using hits@1 values in AvsA graph alignments using all the information from the source database (A), and different data subdomains: crm sequences (B), relations between crm and genes (C), relations between crm and phenotypes (D), and relations between crm and transcription factors (E).

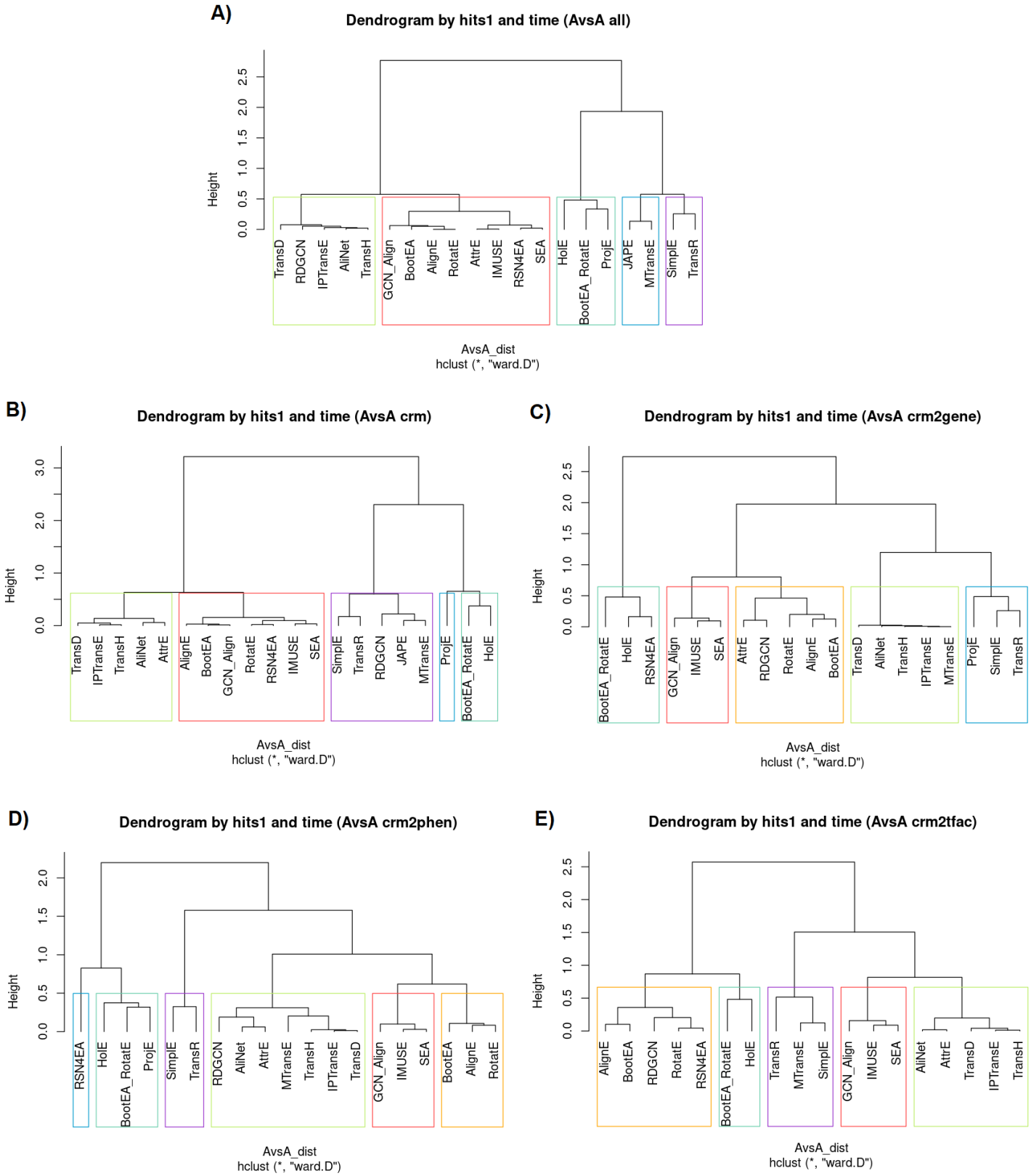
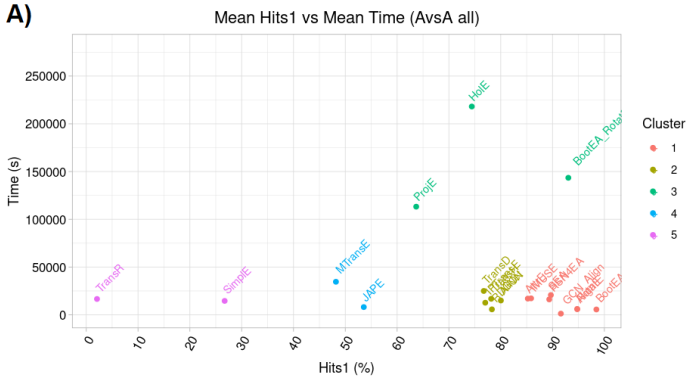
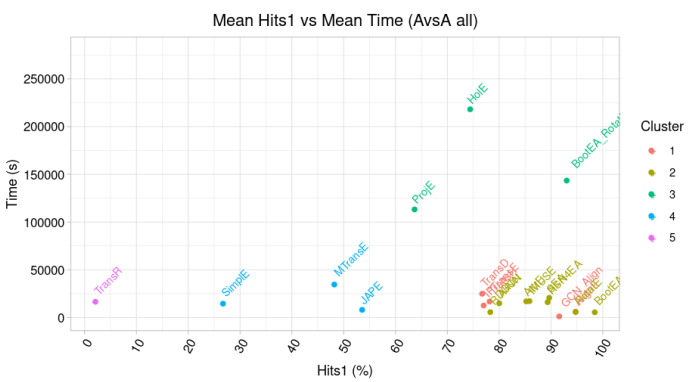


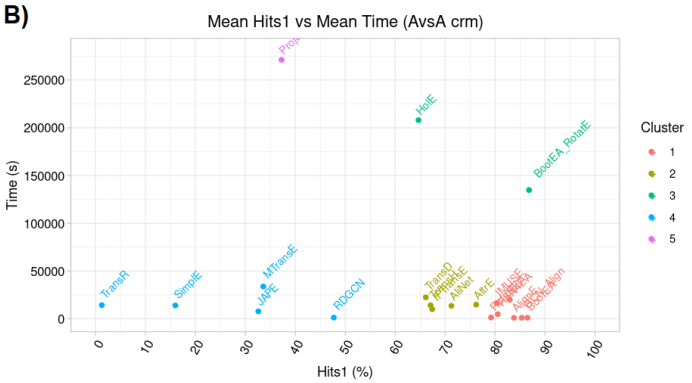
Figure 13. Dendrograms of methods using hits@1 and time values in AvsA alignments with all data domain from the source databases (A), and different data subdomains: crm sequences (B), relations between crm and genes (C), relations between crm and phenotypes (D), and relations between crm and transcription factors (E).

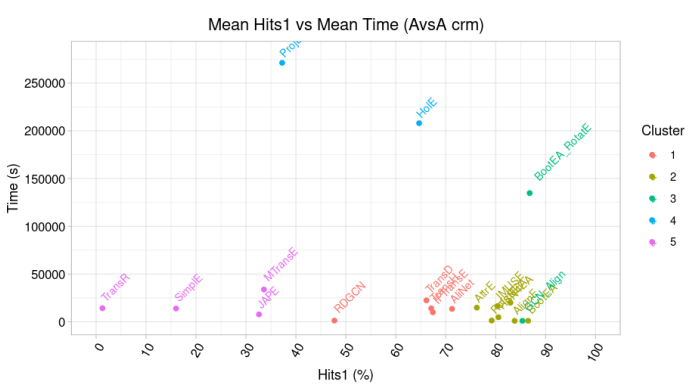
A)


Mean Hits1 vs Mean Time (AvsA all)
Time (s)
Hits1 (%)
Cluster
1
2
3
4
5
TransR
SimplE
MTransE
JAPE
HolE
ProjE
BootEA_RotatE

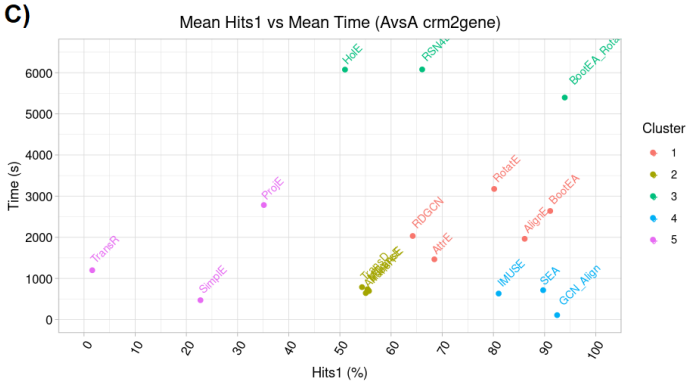

Mean Hits1 vs Mean Time (AvsA all)
Time (s)
Hits1 (%)
Cluster
1
2
3
4
5
TransR
SimplE
MTransE
JAPE
HolE
ProjE
BootEA_RotatE

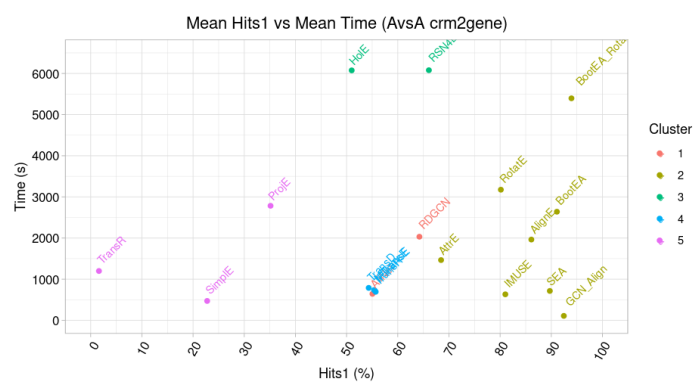
B)


Mean Hits1 vs Mean Time (AvsA crm)
Time (s)
Hits1 (%)
Cluster
1
2
3
4
5
TransR
SimplE
MTransE
JAPE
RDGCN
HolE
ProjE
BootEA_RotatE

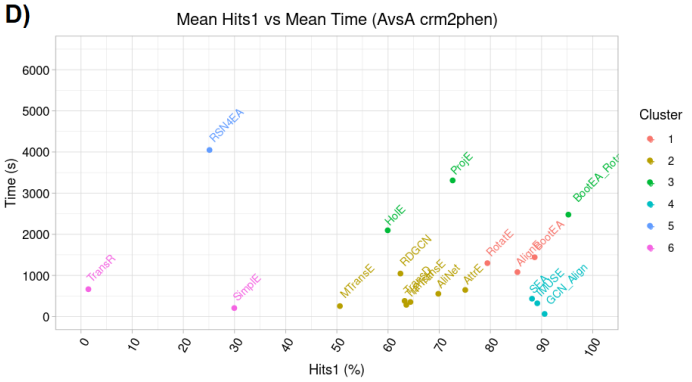

Mean Hits1 vs Mean Time (AvsA crm)
Time (s)
Hits1 (%)
Cluster
1
2
3
4
5
TransR
SimplE
MTransE
JAPE
RDGCN
HolE
ProjE
BootEA_RotatE

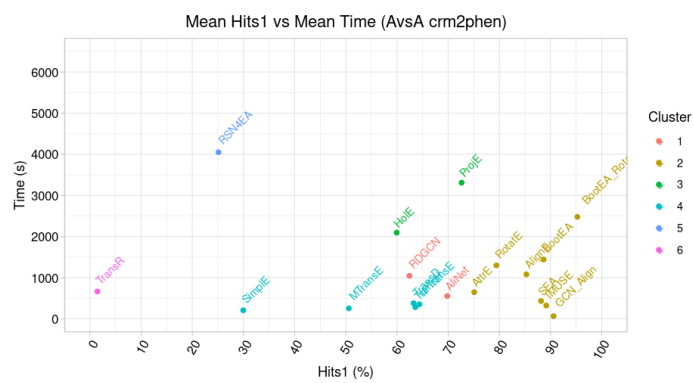
C)


Mean Hits1 vs Mean Time (AvsA crm2gene)
Time (s)
Hits1 (%)
Cluster
1
2
3
4
5
TransR
SimplE
ProjE
HolE
RSNA
RDGCN
AttrE
RotatE
AlignE
BootEA
IMUSE
SEA
GCN_Align


Mean Hits1 vs Mean Time (AvsA crm2gene)
Time (s)
Hits1 (%)
Cluster
1
2
3
4
5
TransR
SimplE
ProjE
HolE
RSNA
RDGCN
AttrE
RotatE
AlignE
BootEA
IMUSE
SEA
GCN_Align

D)


Mean Hits1 vs Mean Time (AvsA crm2phen)
Time (s)
Hits1 (%)
Cluster
1
2
3
4
5
6
TransR
SimplE
RSNAEA
MTransE
ProjE
HolE
RDGCN
AttrE
RotatE
AlignE
BootEA
SEA
IMUSE
GCN_Align


Mean Hits1 vs Mean Time (AvsA crm2phen)
Time (s)
Hits1 (%)
Cluster
1
2
3
4
5
6
TransR
SimplE
RSNAEA
MTransE
ProjE
HolE
RDGCN
AttrE
RotatE
AlignE
BootEA
SEA
IMUSE
GCN_Align

E)

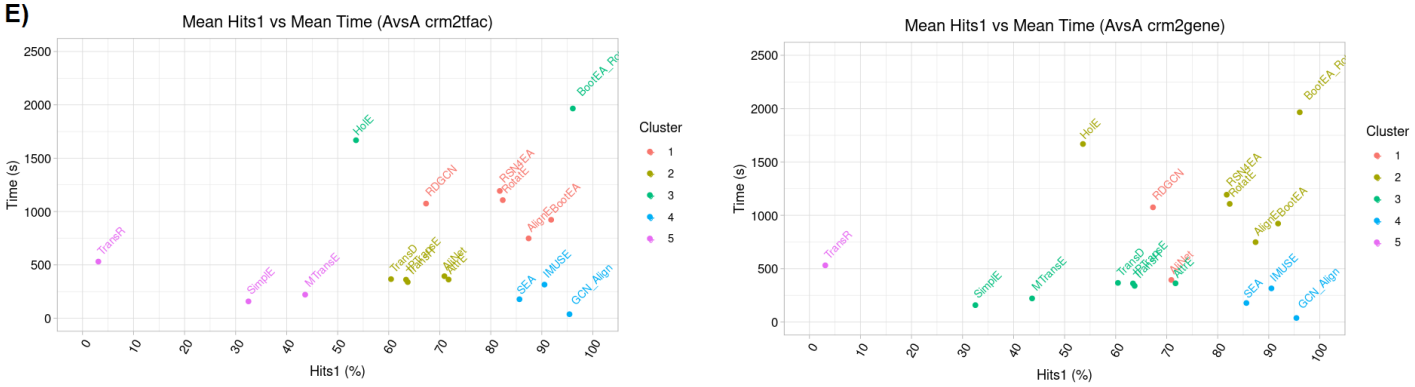
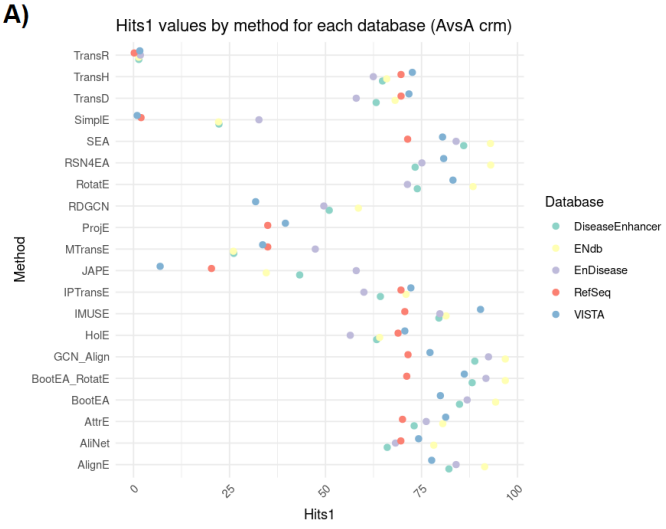


Figure 14. Letf: Plot of methods grouped according to the clusters obtained in the dendrograms by hits@1 and time values (Supplementary Figure 17), AvsA alignments. Right: Plot of methods grouped according to the clusters obtained in the dendrograms including all the alignment metrics (hits@1, hits@5, hits@10, mr, mrr, time). Each row corresponds to a different subdomain of data: all data (A), crm (B), crm2gene (C), crm2phen (D), crm2tfac (E).

A)

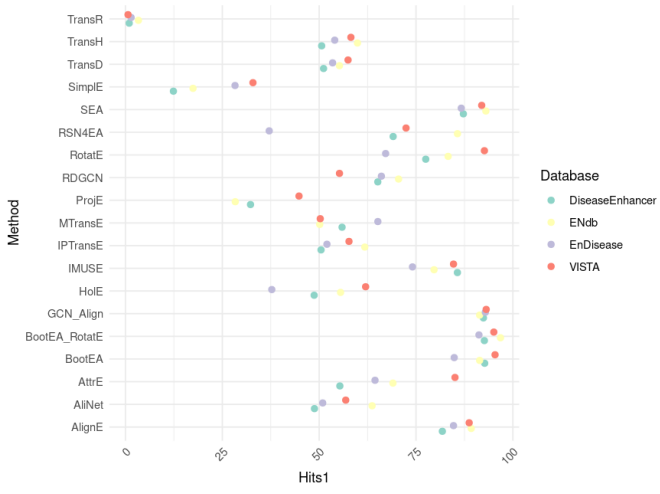


	database	first	second	third	fourth	fifth	weight
1	DiseaseEnhancer	0	4	7	8	1	10.8
2	ENdb	9	3	5	2	1	15.4
3	EnDisease	4	4	5	2	5	12
4	RefSeq	0	5	2	2	11	8.2
5	VISTA	7	4	1	6	2	13.6

	database	first	second	third	fourth	fifth	weight
1	DiseaseEnhancer	0	1	5	2	0	4.6
2	ENdb	6	2	0	0	0	7.6
3	EnDisease	0	4	3	1	0	5.4
4	RefSeq	0	0	0	0	8	1.6
5	VISTA	2	1	0	5	0	4.8

	database	first	second	third	fourth	fifth	weight
1	DiseaseEnhancer	0	1	2	2	0	2.8
2	ENdb	3	2	0	0	0	4.6
3	EnDisease	0	2	3	0	0	3.4
4	RefSeq	0	0	0	0	5	1
5	VISTA	2	0	0	3	0	3.2

B)

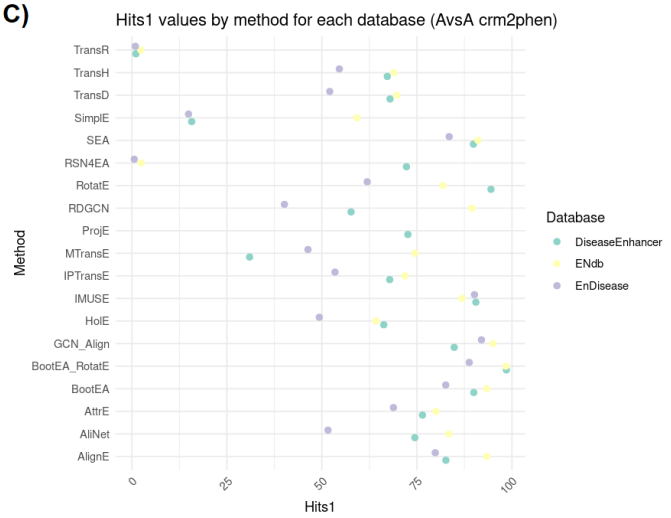


	database	first	second	third	fourth	weight
1	DiseaseEnhancer	1	3	8	7	9
2	ENdb	9	4	4	2	14.5
3	EnDisease	1	4	6	8	9
4	VISTA	8	8	1	2	15

	database	first	second	third	fourth	weight
1	DiseaseEnhancer	1	1	4	2	4.25
2	ENdb	3	2	2	1	5.75
3	EnDisease	0	1	2	5	3
4	VISTA	4	4	0	0	7

	database	first	second	third	fourth	weight
1	DiseaseEnhancer	1	0	3	1	2.75
2	ENdb	2	1	1	1	3.5
3	EnDisease	0	1	1	3	2
4	VISTA	2	3	0	0	4.25

C)



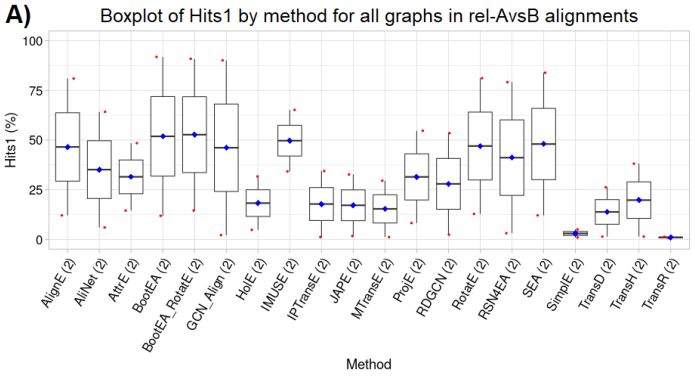
	database	first	second	third	weight
1	DiseaseEnhancer	6	11	2	14
2	ENdb	13	5	1	16.67
3	EnDisease	0	3	16	7.33

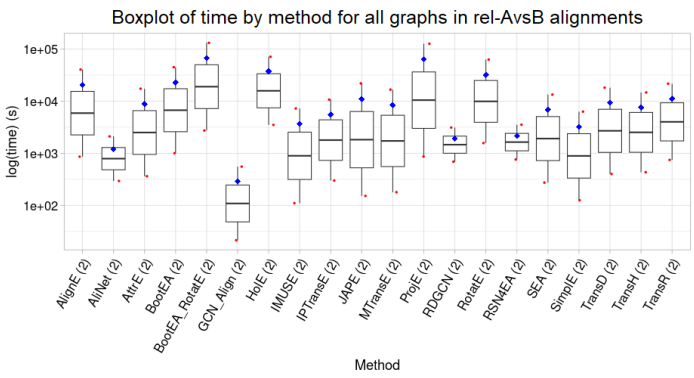
	database	first	second	third	weight
1	DiseaseEnhancer	3	4	1	6
2	ENdb	5	2	1	6.67
3	EnDisease	0	2	6	3.33

	database	first	second	third	weight
1	DiseaseEnhancer	2	2	1	3.67
2	ENdb	3	1	1	4
3	EnDisease	0	2	3	2.33

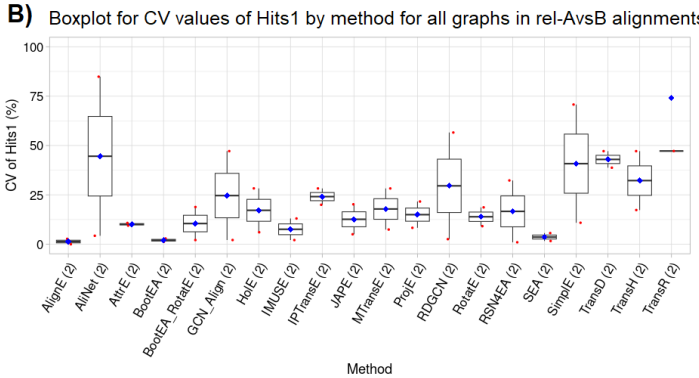
Figure 15. Left: Plot of hits@1 values by method and dataset in AvsA alignments, and considering different data subdomains of the datasets: crm sequences (A), relations crm2gene (B) and crm2phen (C). Right: Count the position of each method in a ranking by hits@1 values and different method combinations: all the methods (first table), eight methods (second table: GCN Align, SEA, IMUSE, BootEA, AlignE, RotatE, AttrE and BootEA-RotatE) and five methods (third table: GCN Align, SEA, IMUSE, AttrE and BootEA-RotatE).

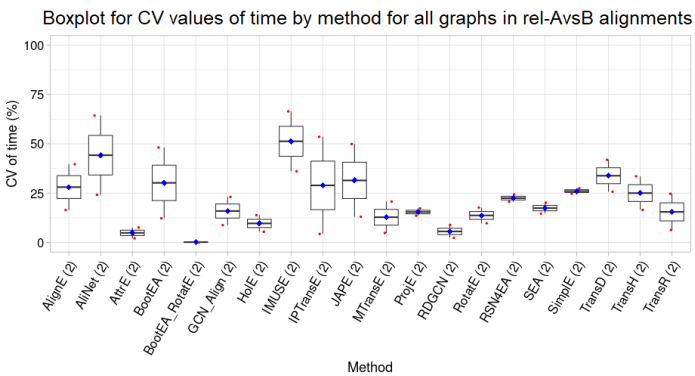
A)





B)





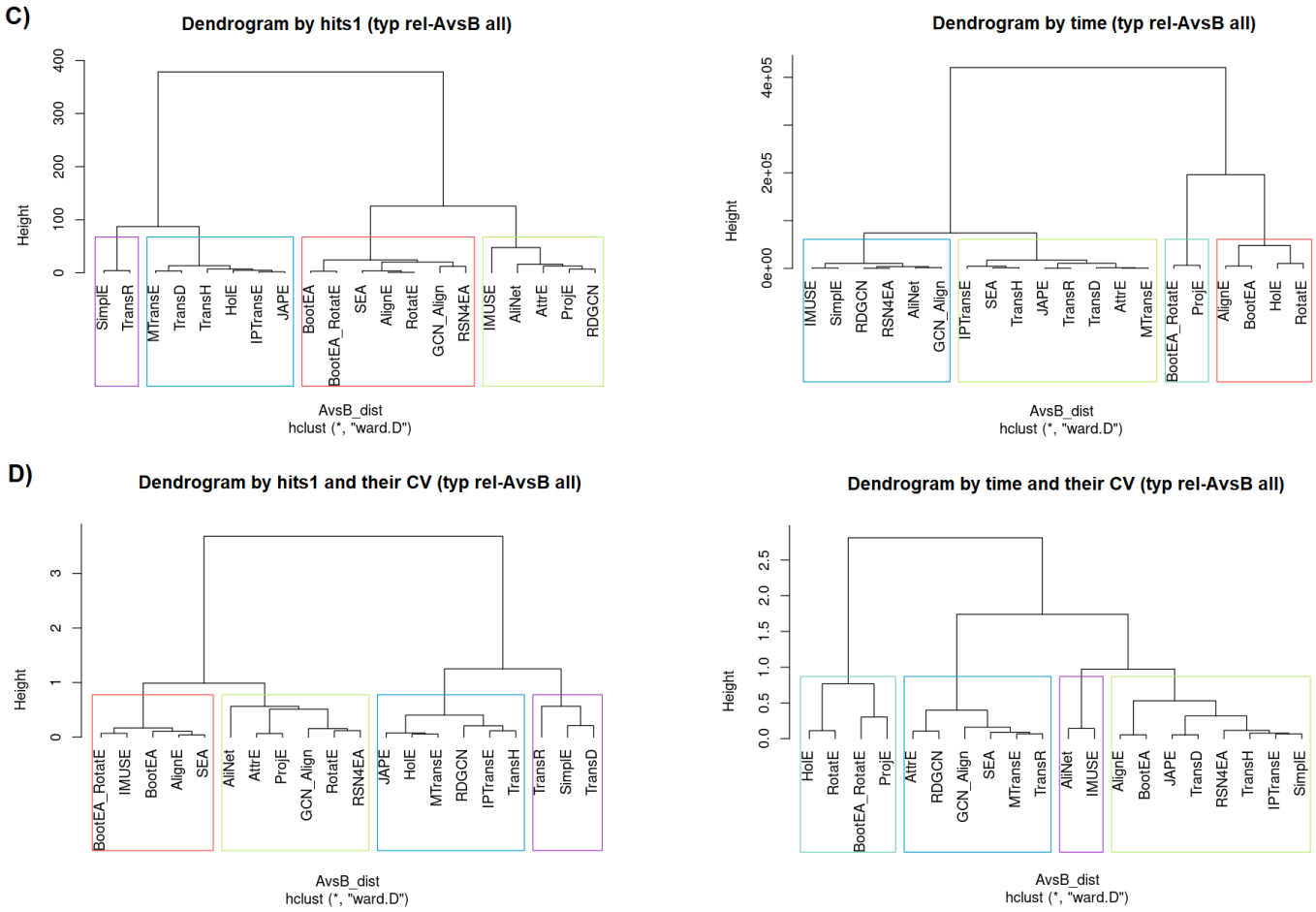


Figure 16. Results corresponding to typical rel-AvsB alignments with all data and using two pairs: EnDisease-DiseaseEnhancer and RefSeq-VISTA. Left: hits@1 values. Right: Time values. A) Boxplot of hits@1 and time values. B) Coefficients of variation of hits@1 and time values. C) Dendrogram of methods with 4 clusters using hits@1 values and time values. D) Alternative dendrograms obtained by adding the coefficients of variation to the hits@1 and time values.

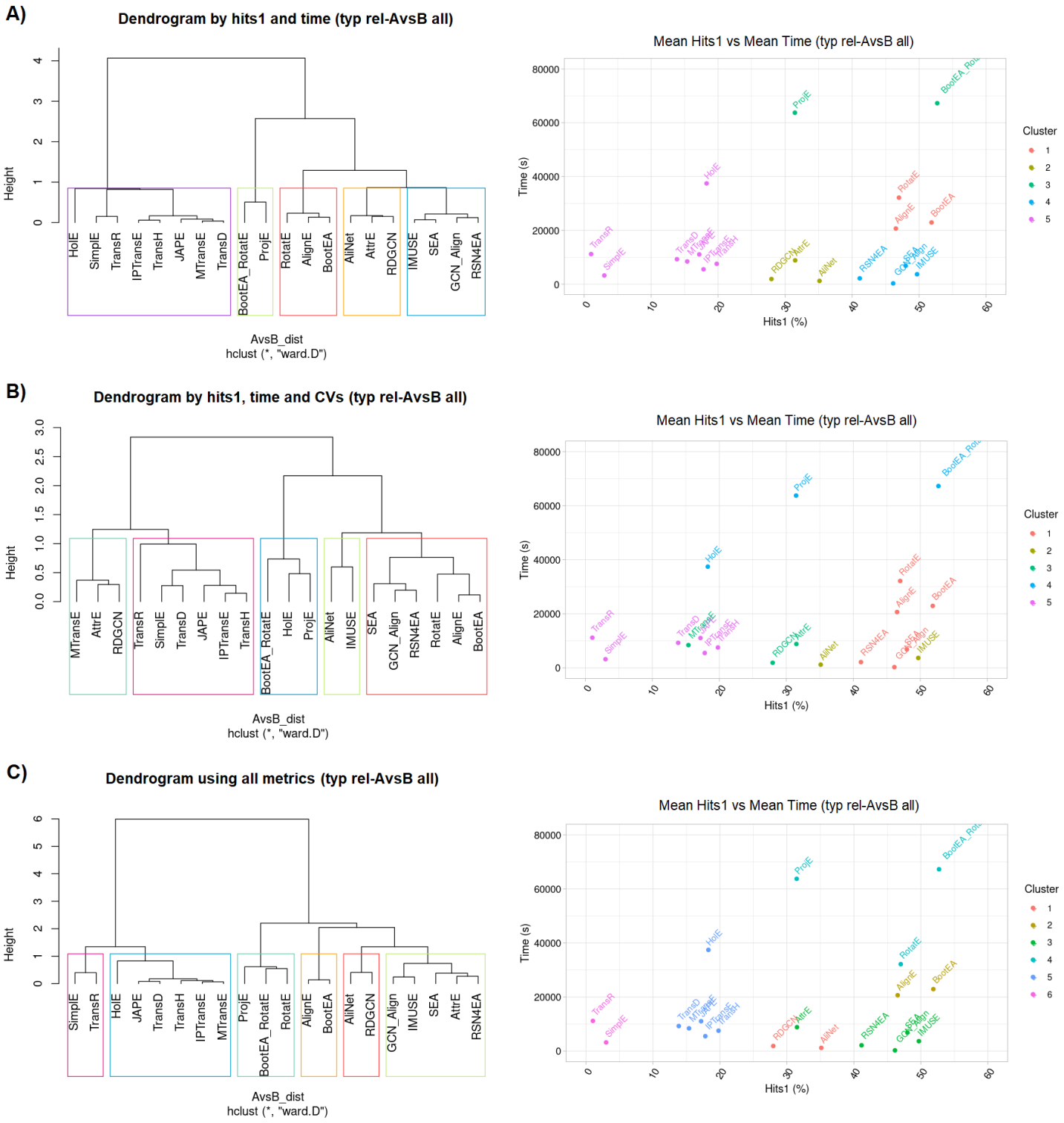


Figure 17. Values associated with typical rel-AvsB alignments using the pairs EnDisease-DiseaseEnhancer, and RefSeq-VISTA. Clustering of methods using: hits@1 and time values (A); hits@1, time and their coefficients of variation (B); and using all metrics (hits@1, hits@5, hits@10, mr, mrr, time) (C).

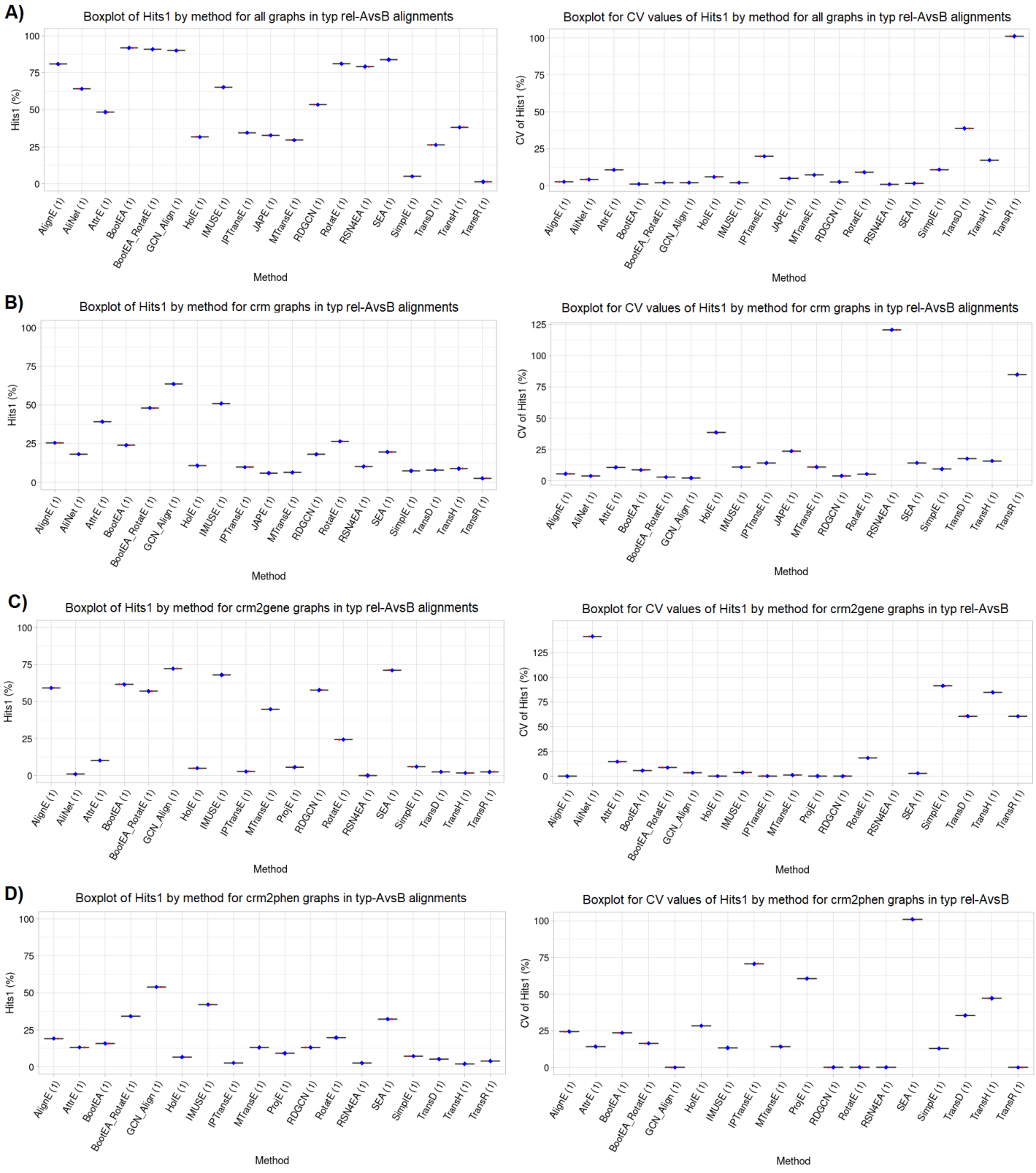


Figure 18. Boxplot for hits@1 values (left) and their coefficients of variation (right) in typical rel-AvsB alignments by subgraphs. The EnDisease-DiseaseEnhancer pair was used, which is the pair that has different subgraphs and enough common entities to train a model. Values using all data (A) and the different subgraphs (B: crm, C: crm2gene, D: crm2phen).

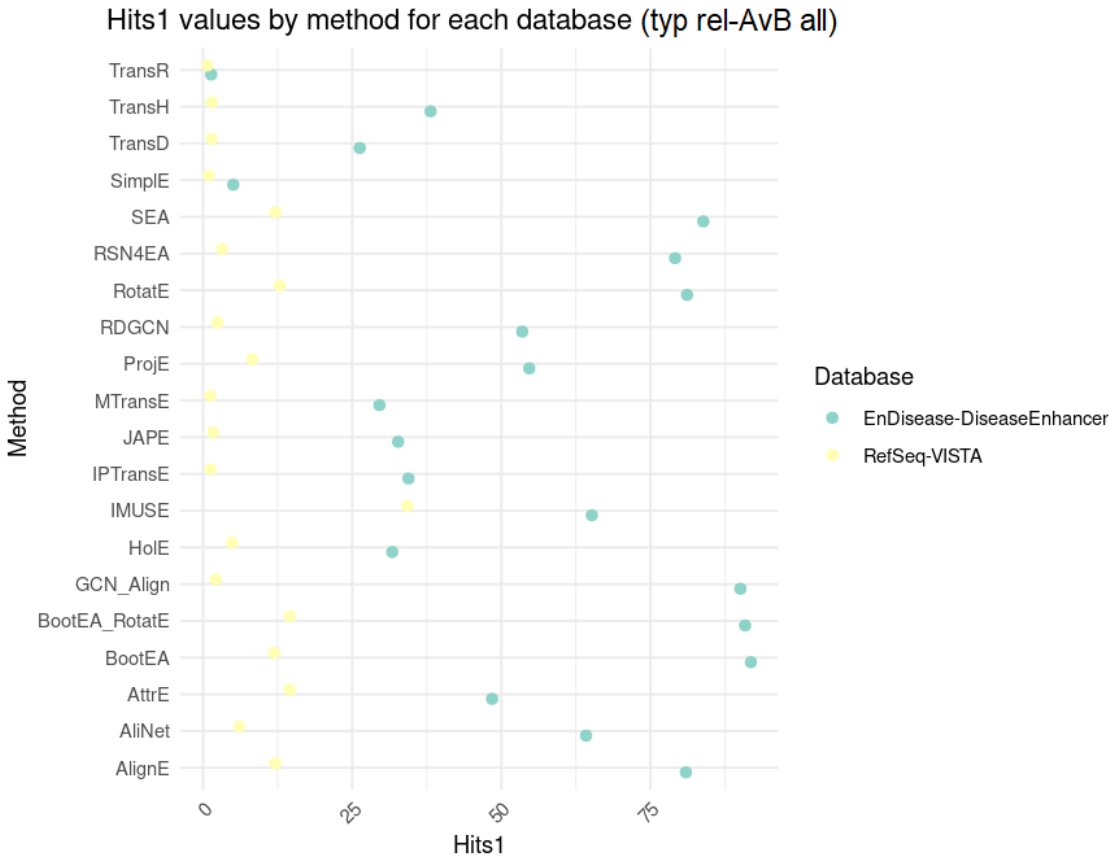


Figure 19. Plot of hits@1 values by method and dataset in typical rel-AvsB alignments with all data domains.

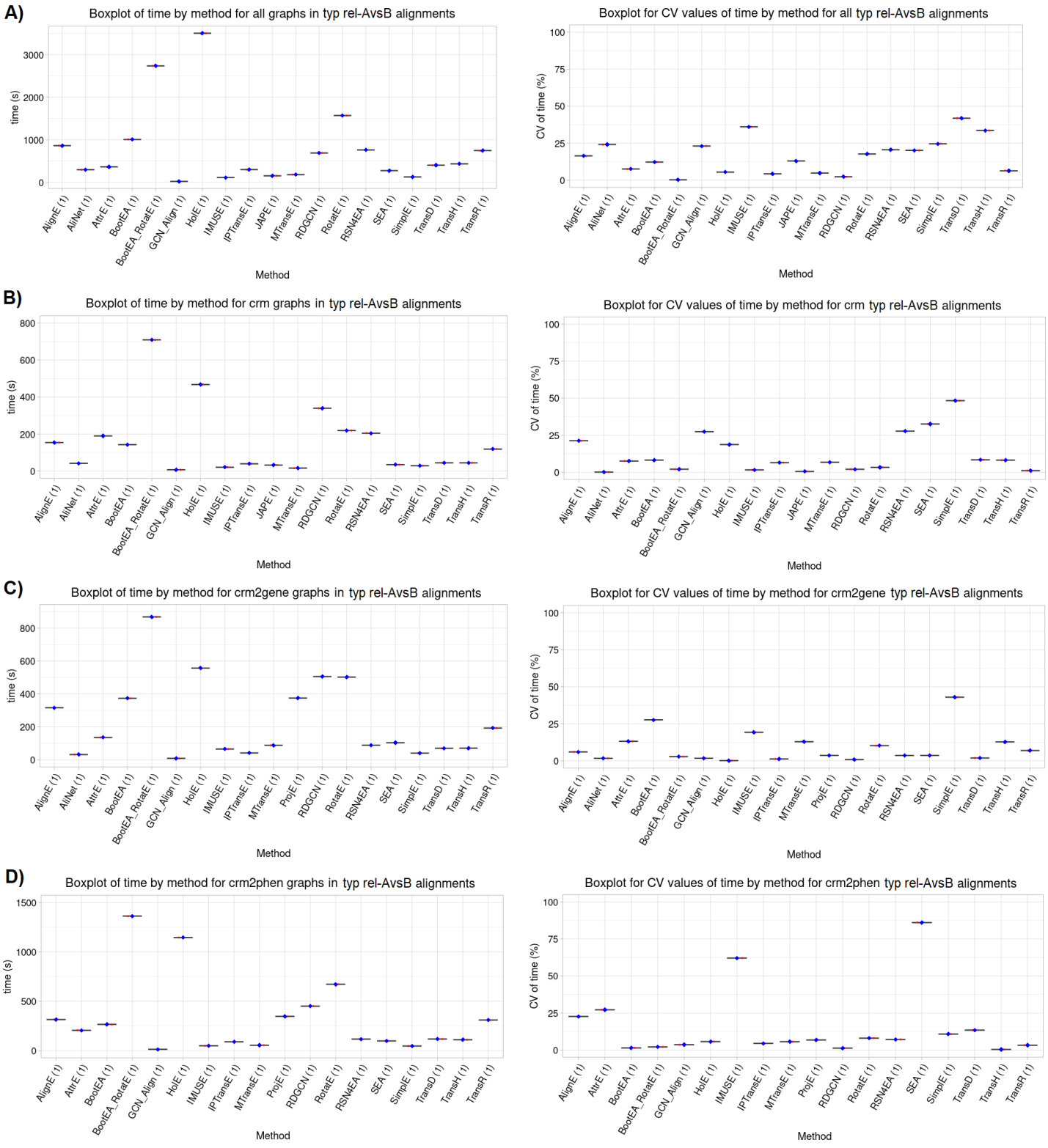


Figure 20. Plot for time values (left) and their coefficients of variation (right) in typical rel-AvsB alignments using different subgraphs. The EnDisease-DiseaseEnhancer pair was used, which is the pair that has different subgraphs and enough common entities to train a model. Values using all data (A) and the different subgraphs (B: crm, C: crm2gene, D: crm2phen).

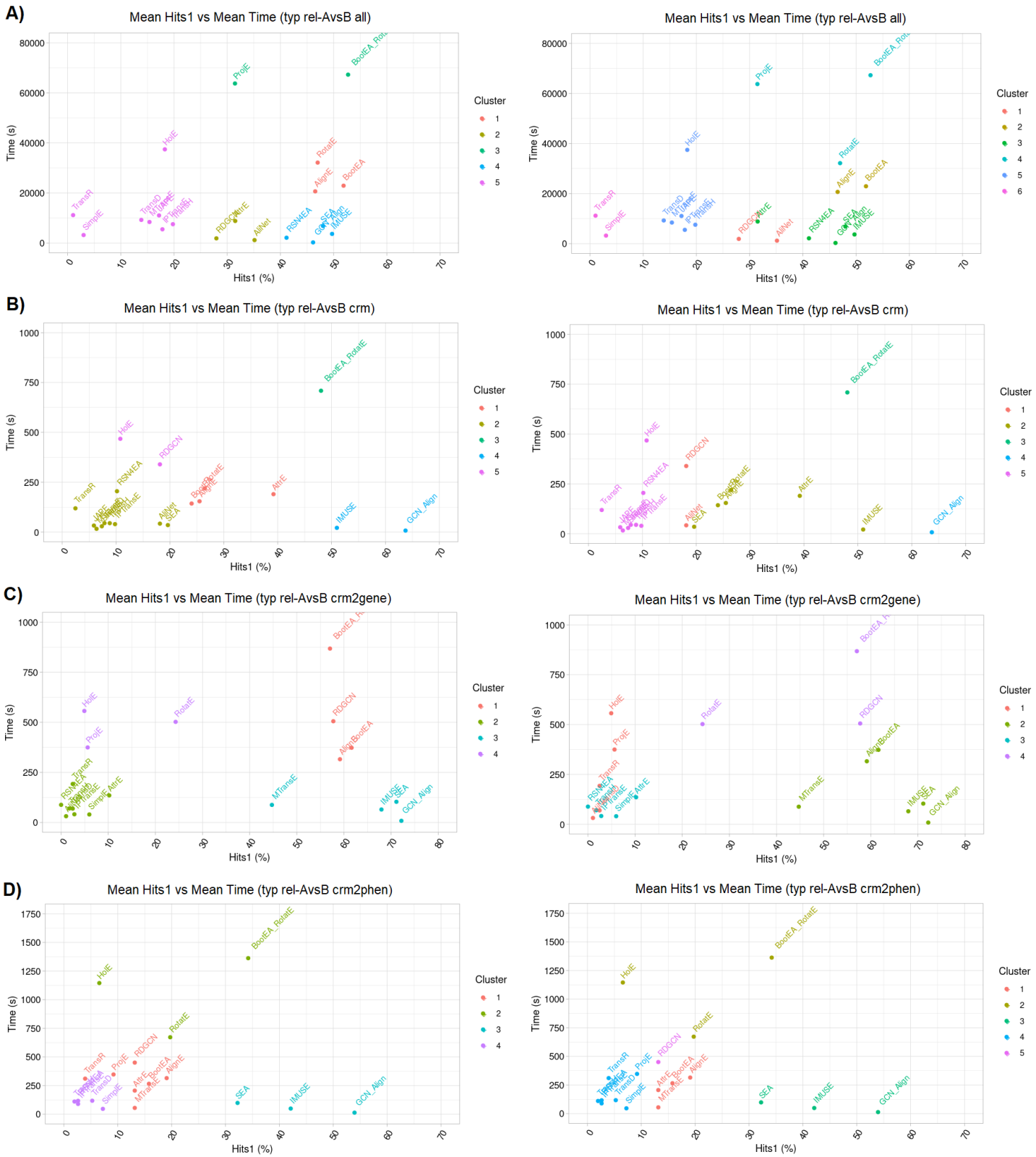


Figure 21. Values from typical rel-AvsB alignments with different domains. Letf: Plot of methods grouped according to the clusters obtained in the dendrograms by hits@1 and time values. Right: Plot of methods grouped according to the clusters obtained in the dendrograms including all the alignment metrics. Each row corresponds to a different subdomain of data: all data (A), crm (B), crm2gene (C), crm2phen (D). Note that the results of A correspond to the average obtained from two different alignments: EnDisease-DiseaseEnhancer and RefSeq-VISTA. On the contrary, B-D correspond to the results of EnDisease-DiseaseEnhancer.

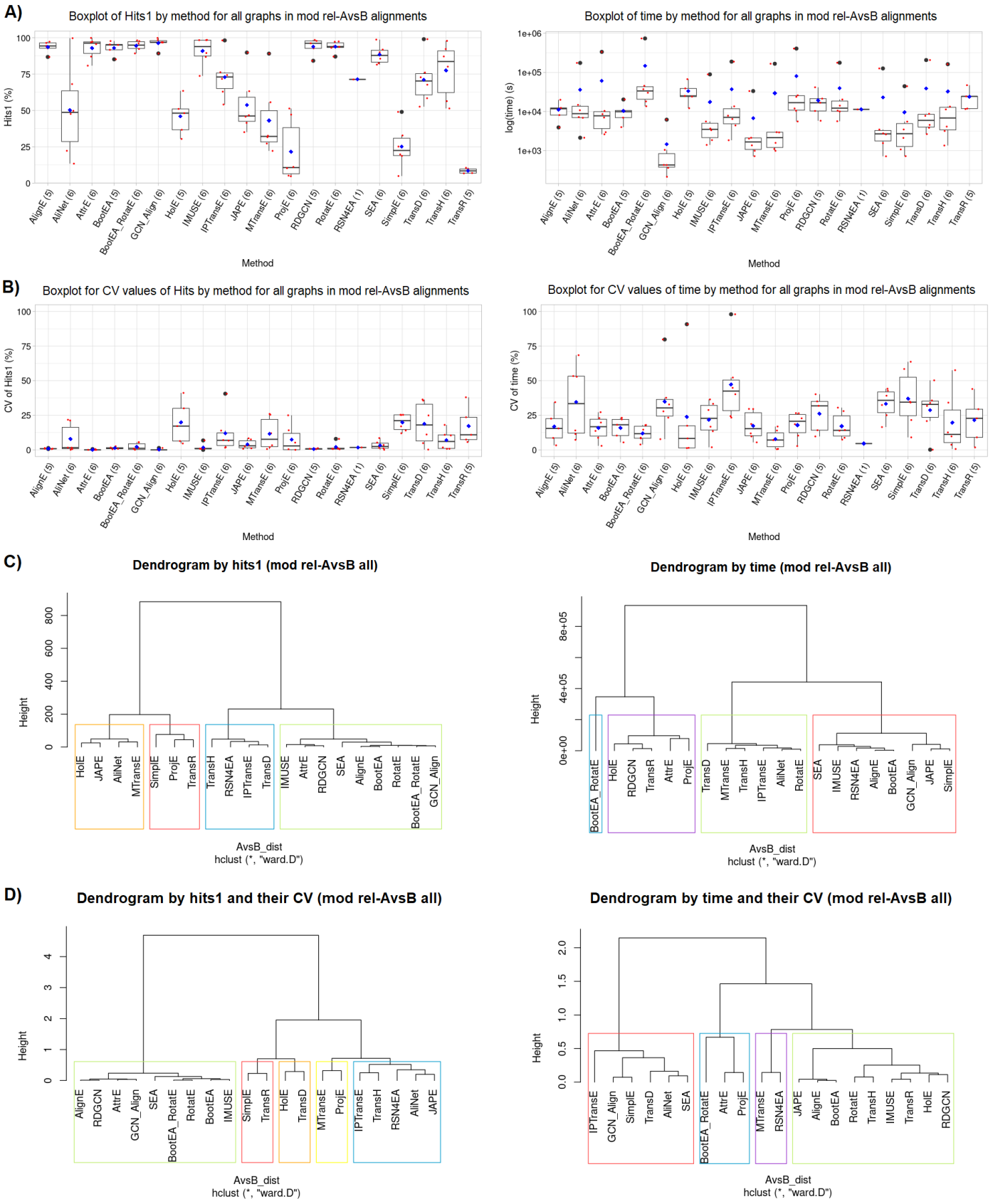


Figure 22. Results corresponding to modified rel-AvsB alignments with all data (A+B vs B+A). Left: hits@1 values. Right: Time values. A) Boxplot of hits@1, and time values. B) Boxplot of coefficients of variation of hits@1, and time values. C) Dendrogram of methods with 4 clusters using hits@1 values, and time values. D) Alternative dendrograms obtained by adding the coefficients of variation to the hits@1, and time values.

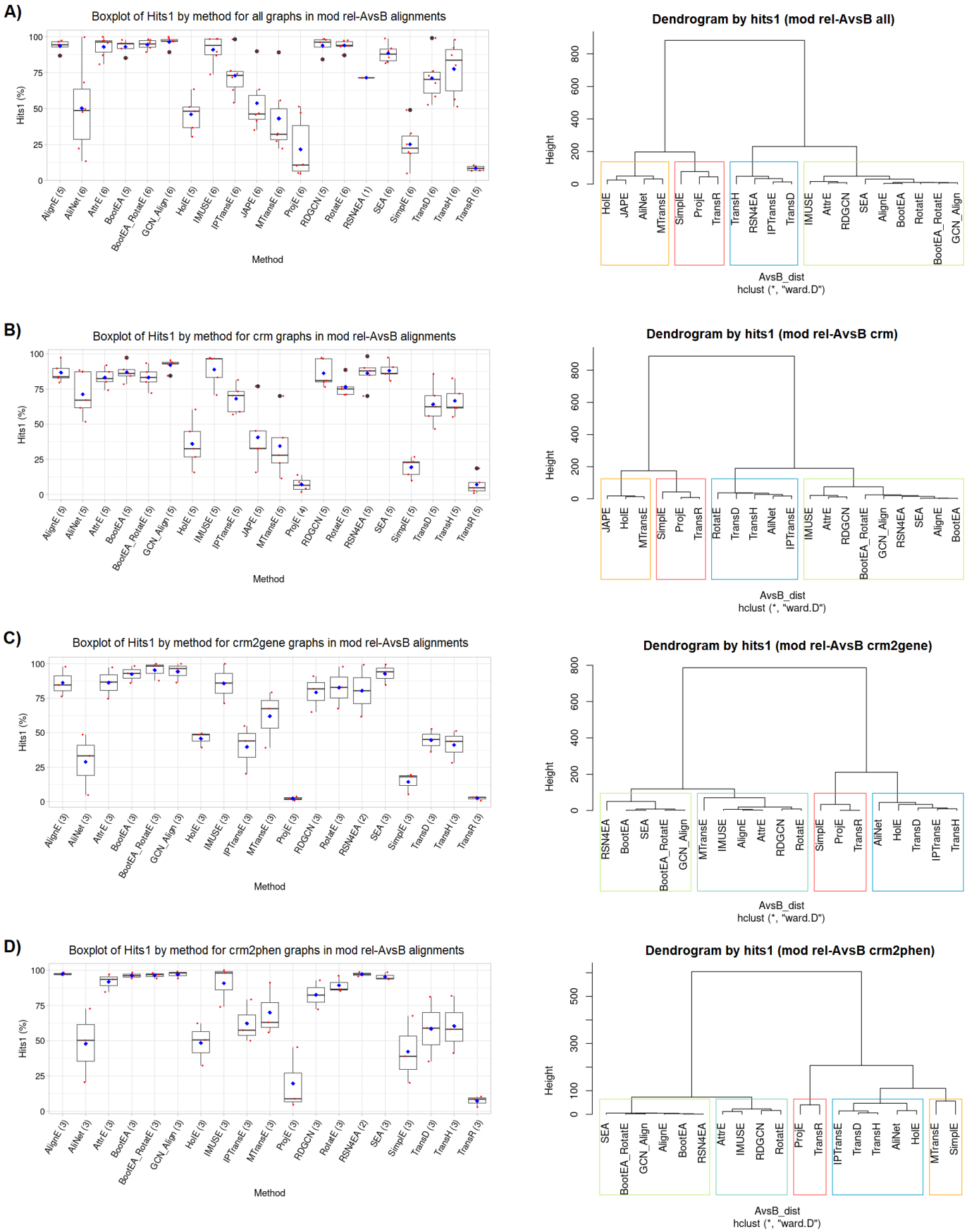


Figure 23. Boxplot and dendrograms in modified rel-AvsB alignments using hits@1 values and different data subdomains: all (A), crm (B), crm2gene (C) and crm2phen (D).

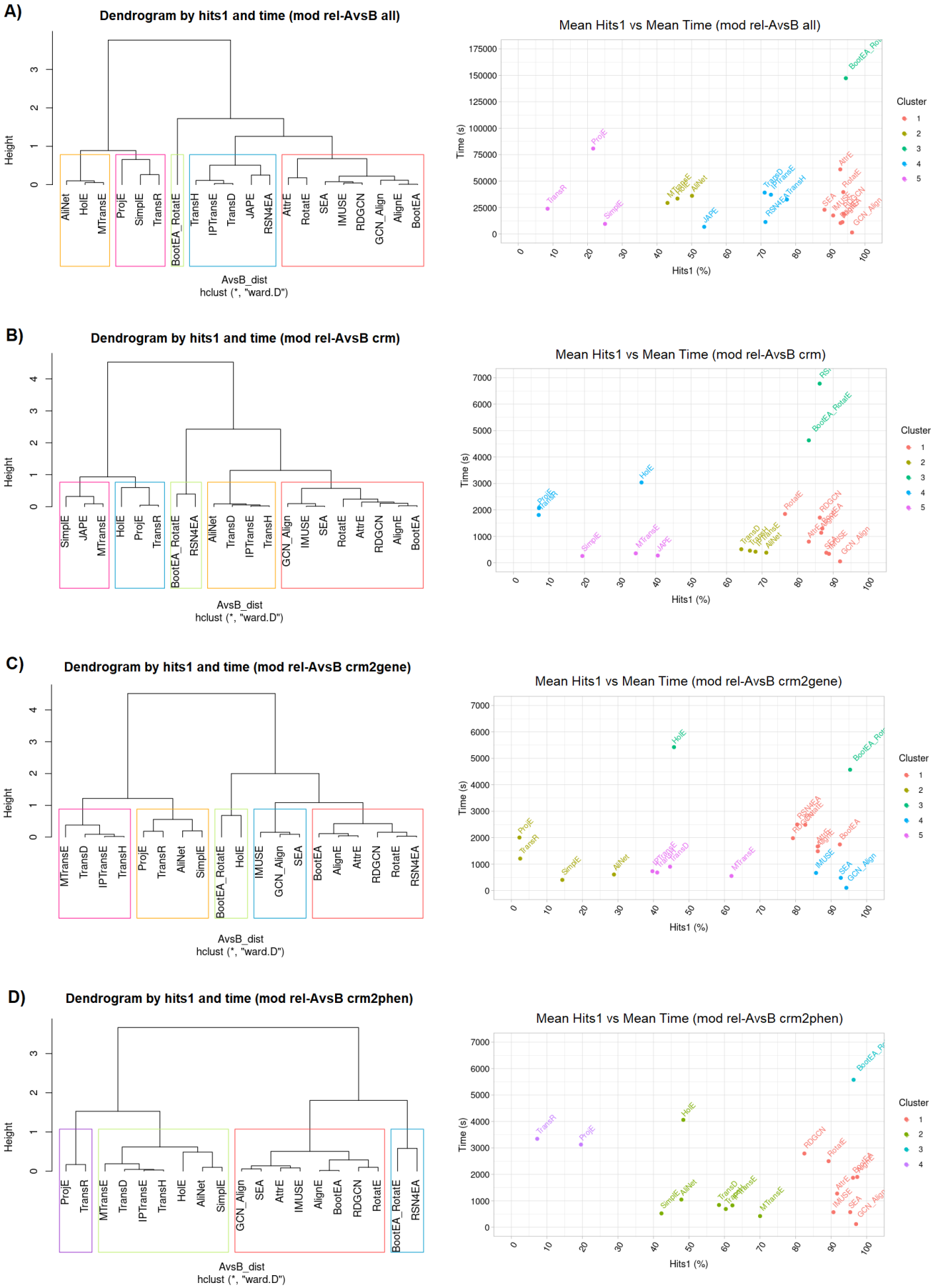


Figure 24. Clustering methods by hits@1 and time in mod rel-AvsB using different data domains: all data (A), crm (B), crm2gene (C), and crm2phen (D).

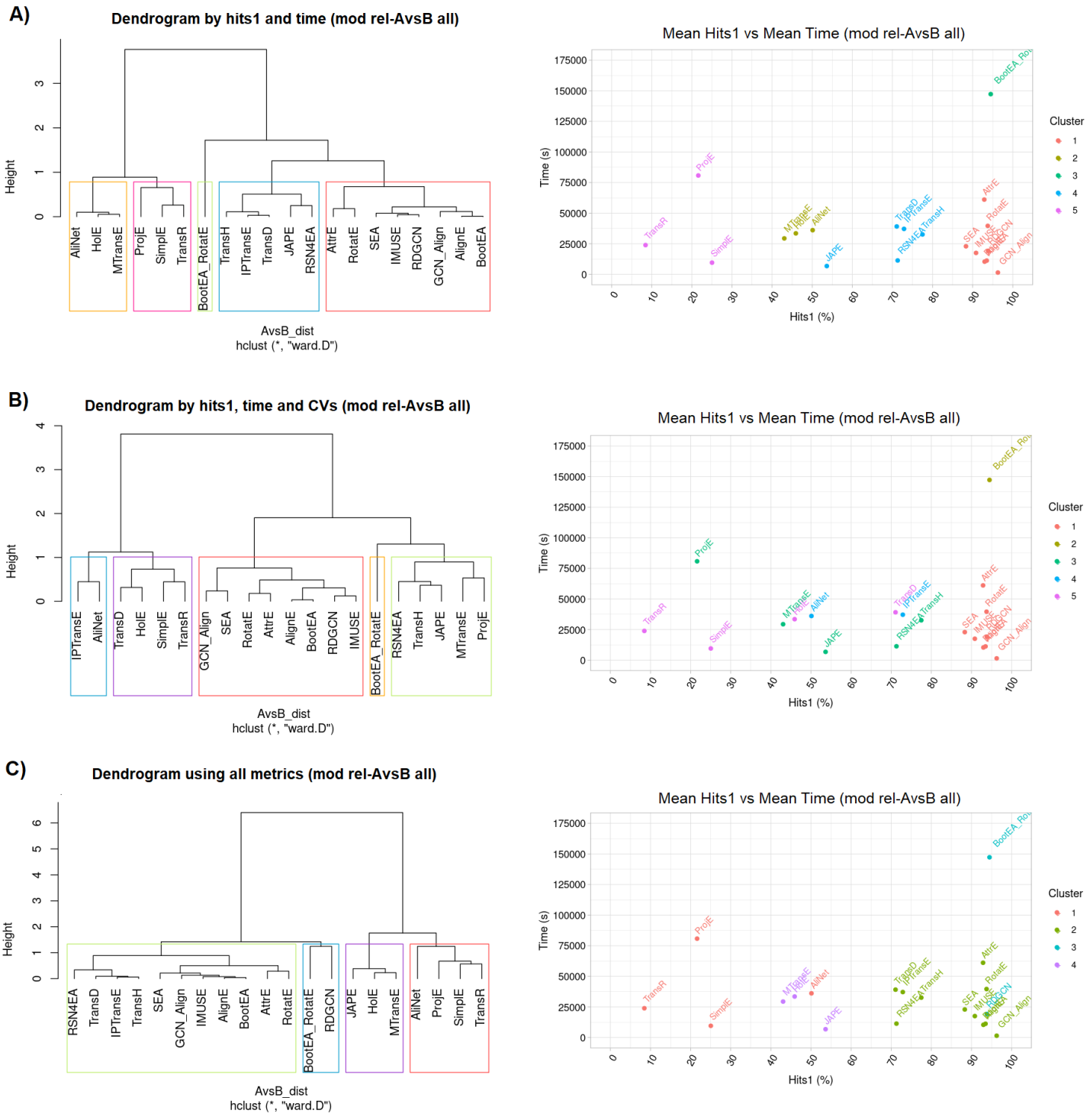


Figure 25. Clustering of methods, in modified rel-AvsB alignments, by hits@1 and time values (A), and including CV values (B), and considering all alignment metrics (C).

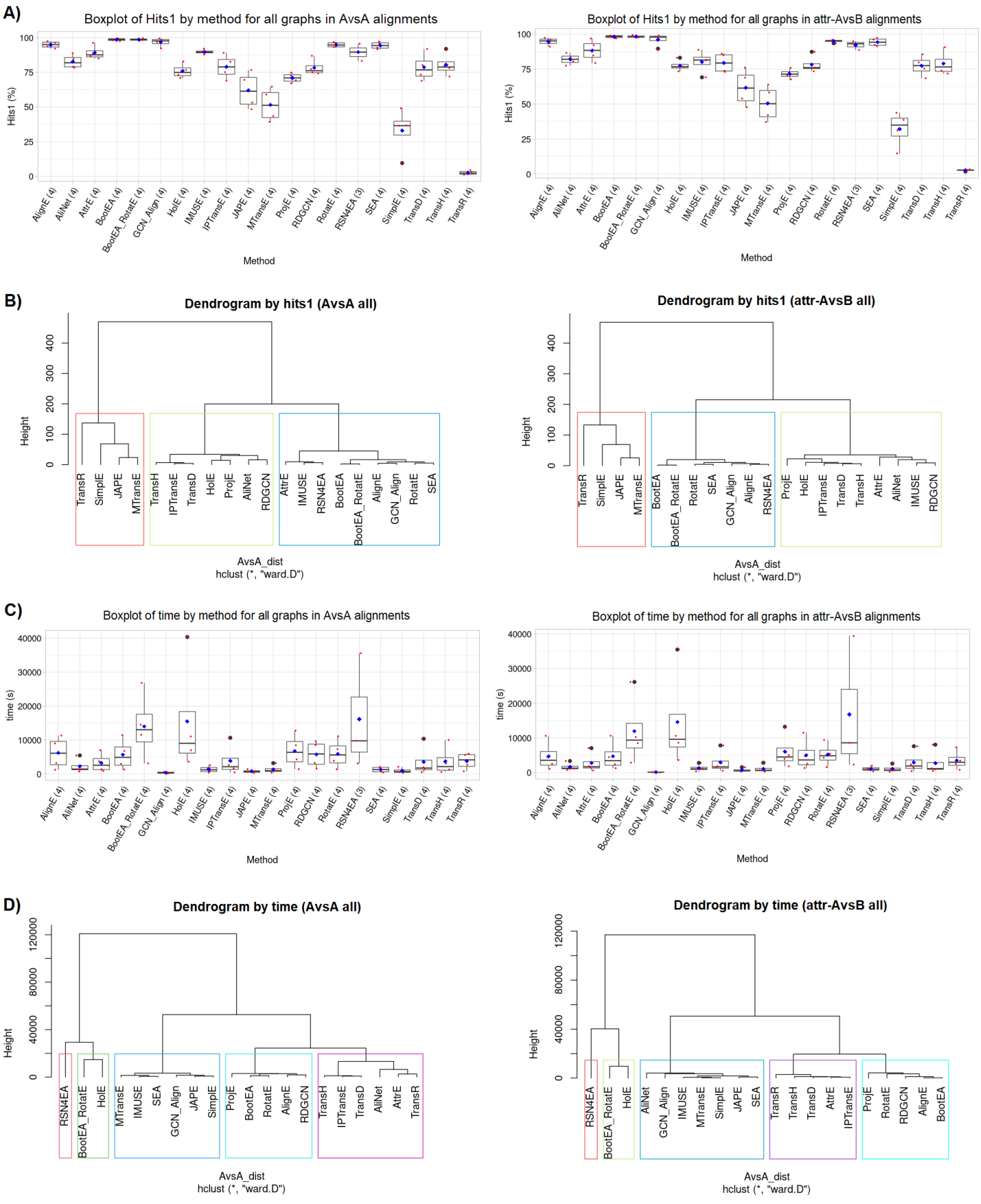


Figure 26. Boxplot and comparative dendrograms of hits@1 and time values between alignment results of AvsA (left) and attr-AvsB (right) type. All data domains were used.

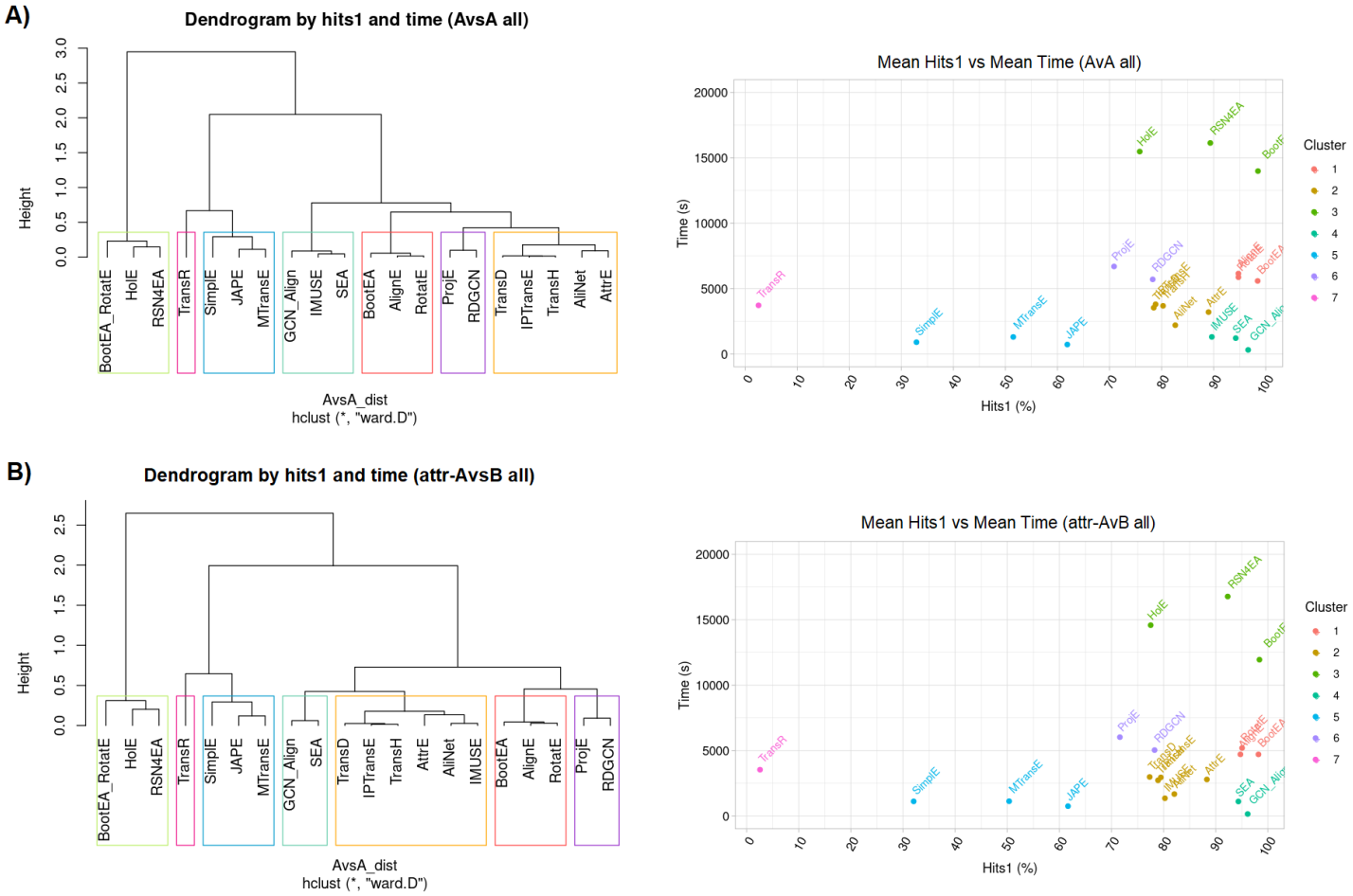


Figure 27. Comparison of dendrograms and plots of methods using hits@1 and time values. A) AvsA alignments. B) attr-AvsB alignments. All data domains were used.

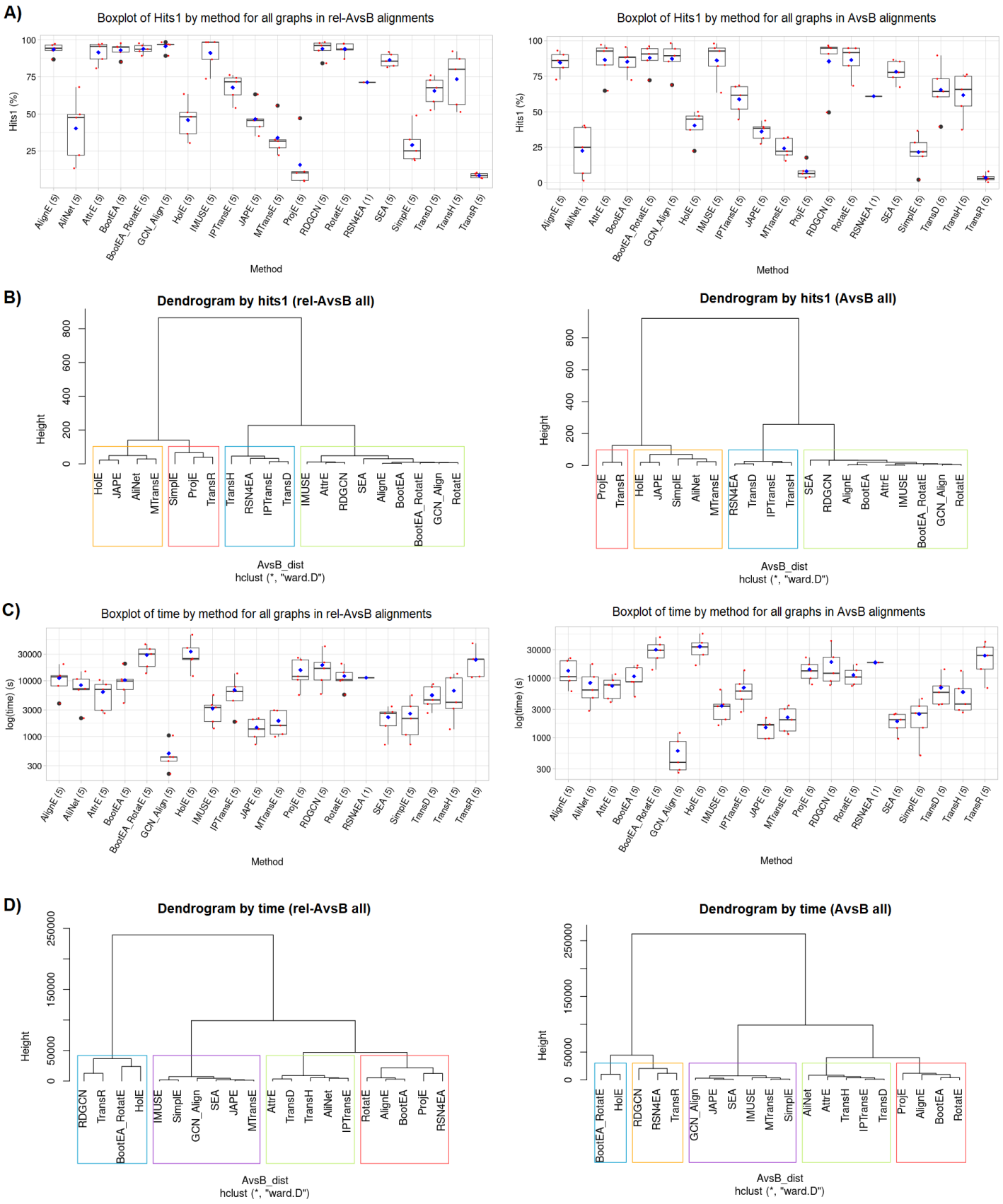


Figure 28. Boxplot and comparative dendrograms of hits@1 and time values between alignment results of modified rel-AvsB (left) and modified AvsB (right) alignments. All data domains were used.

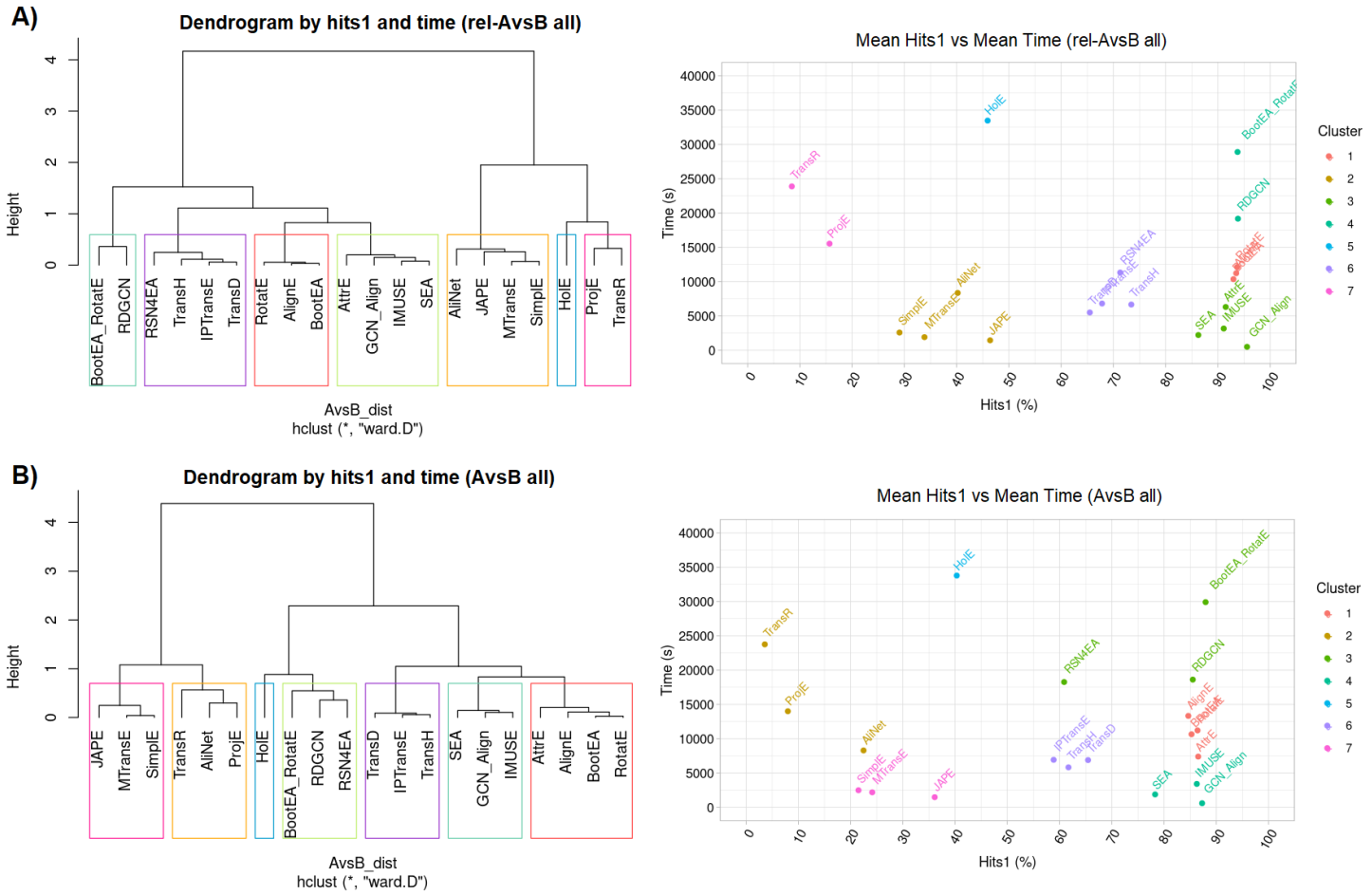


Figure 29. Comparison of dendrograms and plots of methods using hits@1 and time values. A) Modified rel-AvsB alignments. B) Modified AvsB alignments. All data domains were used.