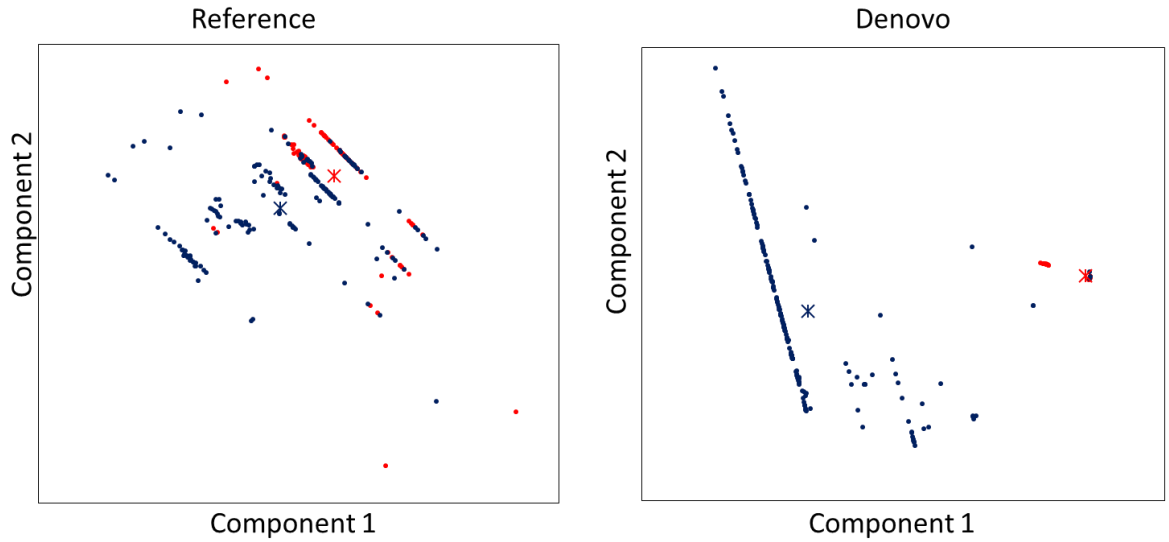
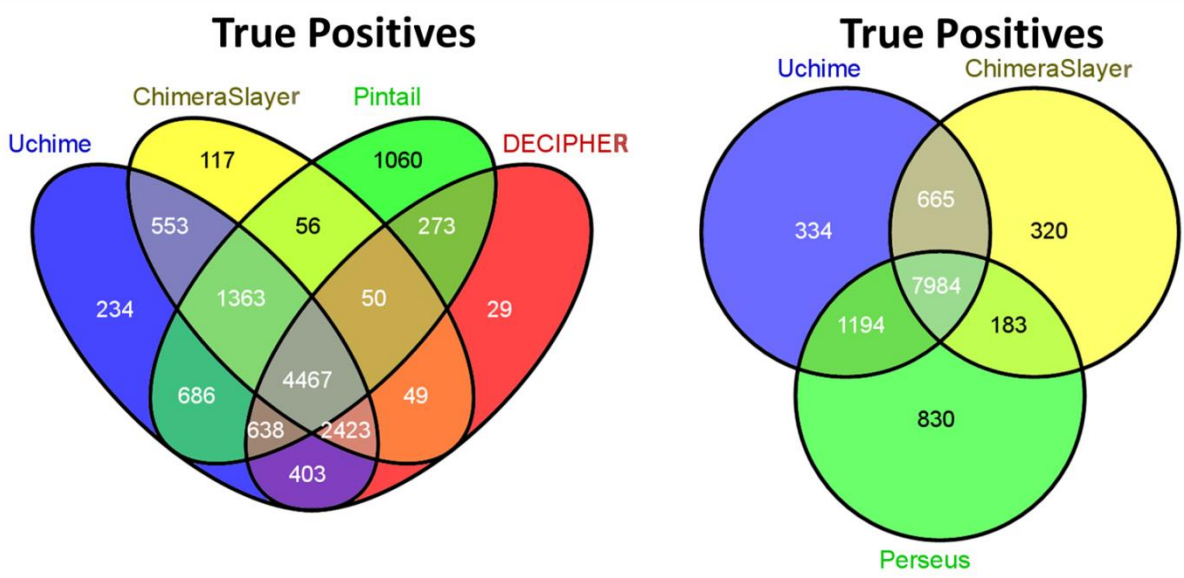


Figure S1. *Plot illustrating the principle component analysis on the training data for both the reference and* de novo *application, where principal component 1 is in the X axis, and principal component 2 in the Y-axis. The red points represent the non-chimeric class and the blue points represent the chimeric class, respectively. The red asterisk represents the centroid of the non-chimeric class, and the blue asterisk represent the centroid of the chimeric class.*



FigureS2*: Venn diagrams showing the overlap between the chimeric sequences correctly predicted by individual chimera detection tools in the Mock2-a dataset. Left: reference based tools; right:* de novo *tools.*

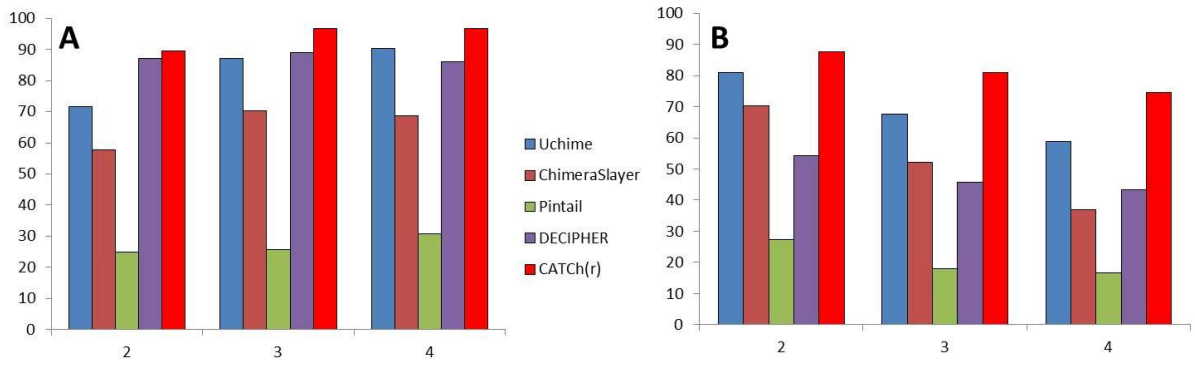


Figure S3*: Sensitivity (in percentage) of all five reference-based algorithms (UCHIME, ChimeraSlayer, Pintail, DECIPHER and CATCh) for detection of different types of chimeras, i.e. having two (bimera), three (trimera) and four (tetramera) parents, applied on (A) Simu3 and (B) Simu2.*

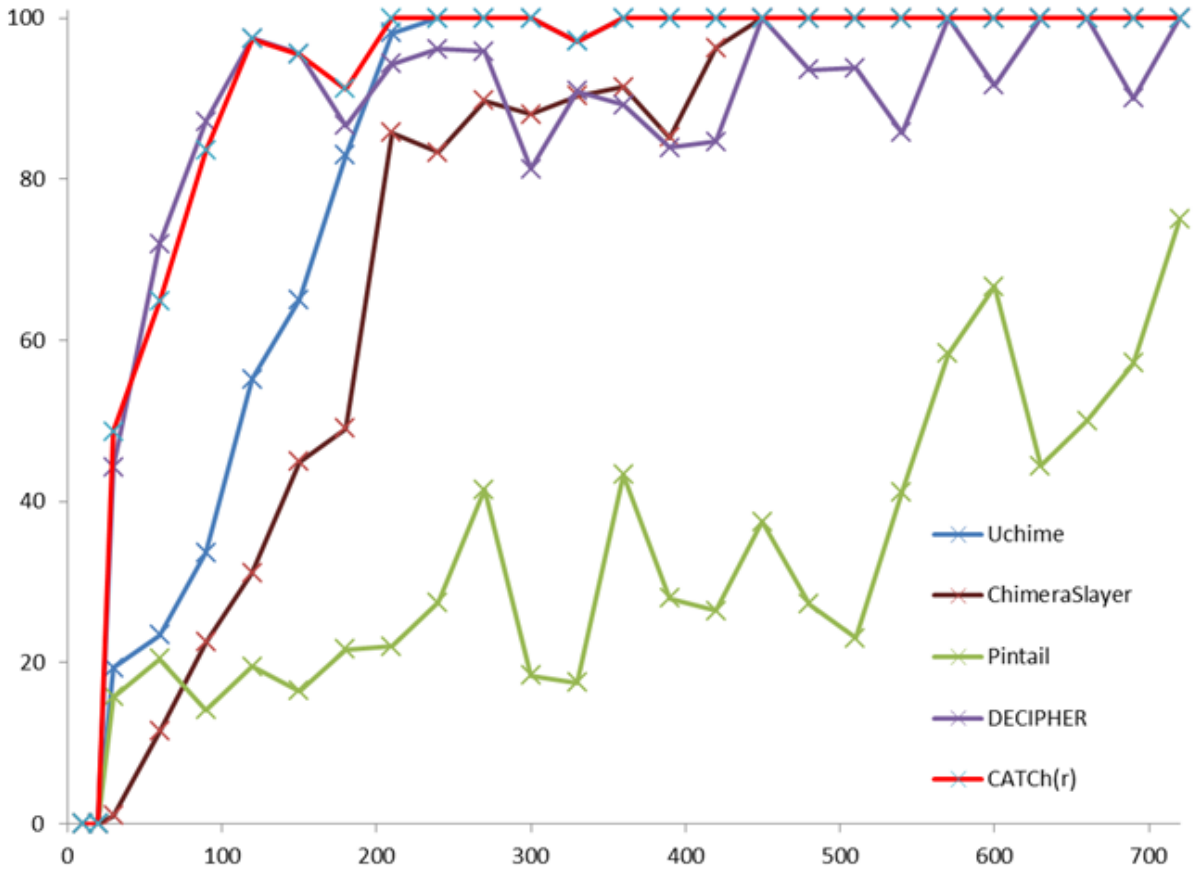


Figure S4: *Effect of the chimeric range (X-axis) on the sensitivity of different algorithms (Y-axis) (UCHIME, ChimeraSlayer, Pintail, DECIPHER and CATCh), using Simu3 subset (b).*

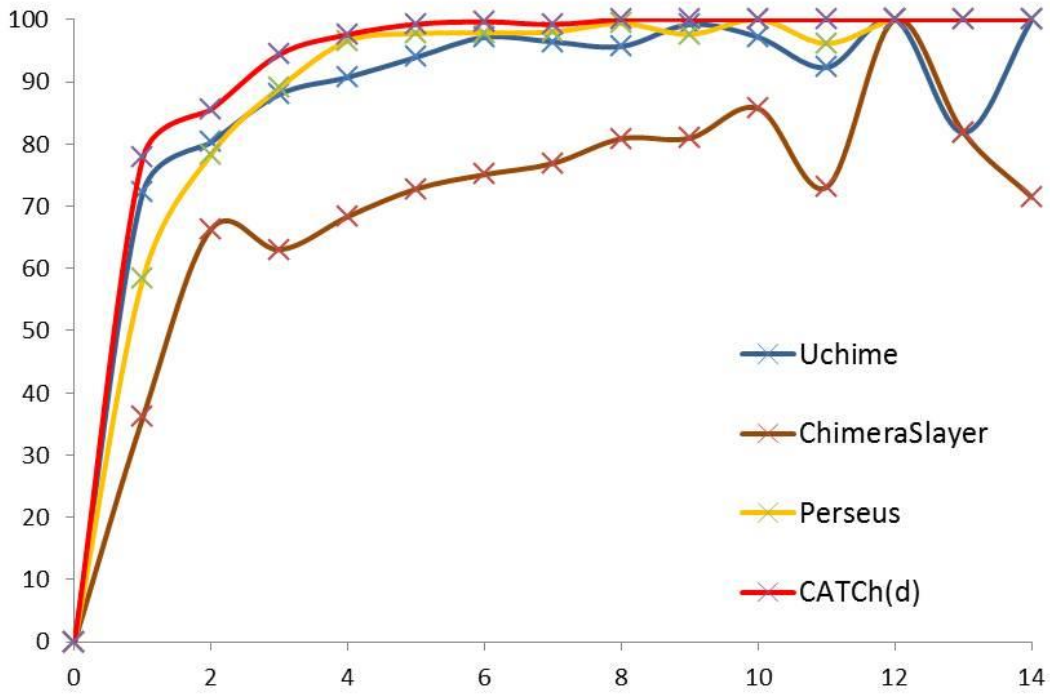


Figure S5*: Illustration of the effect of the divergence of the chimeric sequences (X-axis) on sensitivity of the different* de novo *chimera detection tools (y-axis) using the Mock1 dataset with a divergence ranging from 1 to 14%. As shown, UCHIME* de novo*, ChimeraSlayer* de novo*, Perseus and CATCh* de novo *tend to detect chimeras with higher divergence more accurately than chimeras with lower divergence. As shown in the figure, CATCh produced the best performance compared to the other tools.*

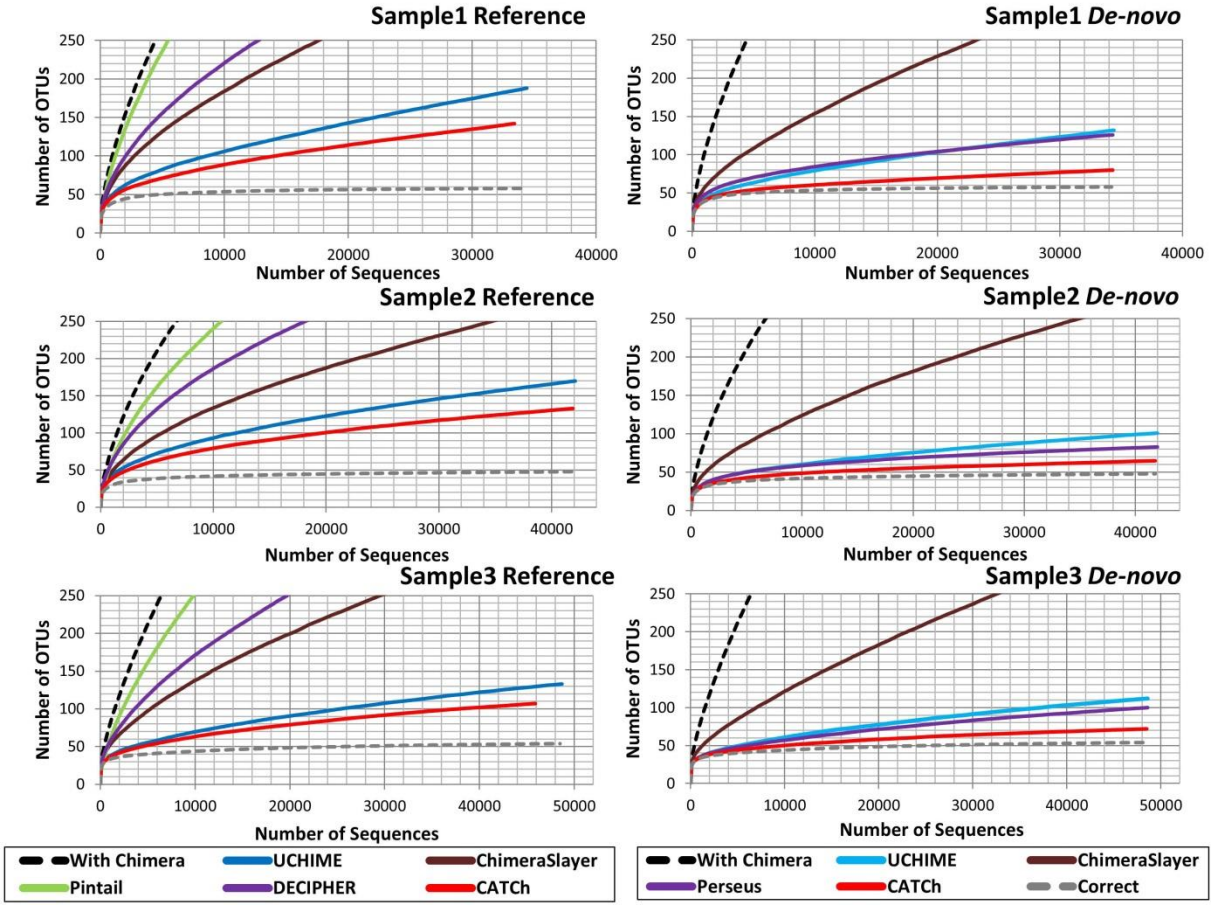


Figure S6: *Rarefaction curves for Mock1 (sample 1,2 and 3) including the datasets i) without any chimera removal ('with chimera'), ii) removing all chimeric sequences ('correct'), and iii) with chimera removal using all tested chimera detection tools (reference-based approaches on the left side and de novo approaches on the right side).*

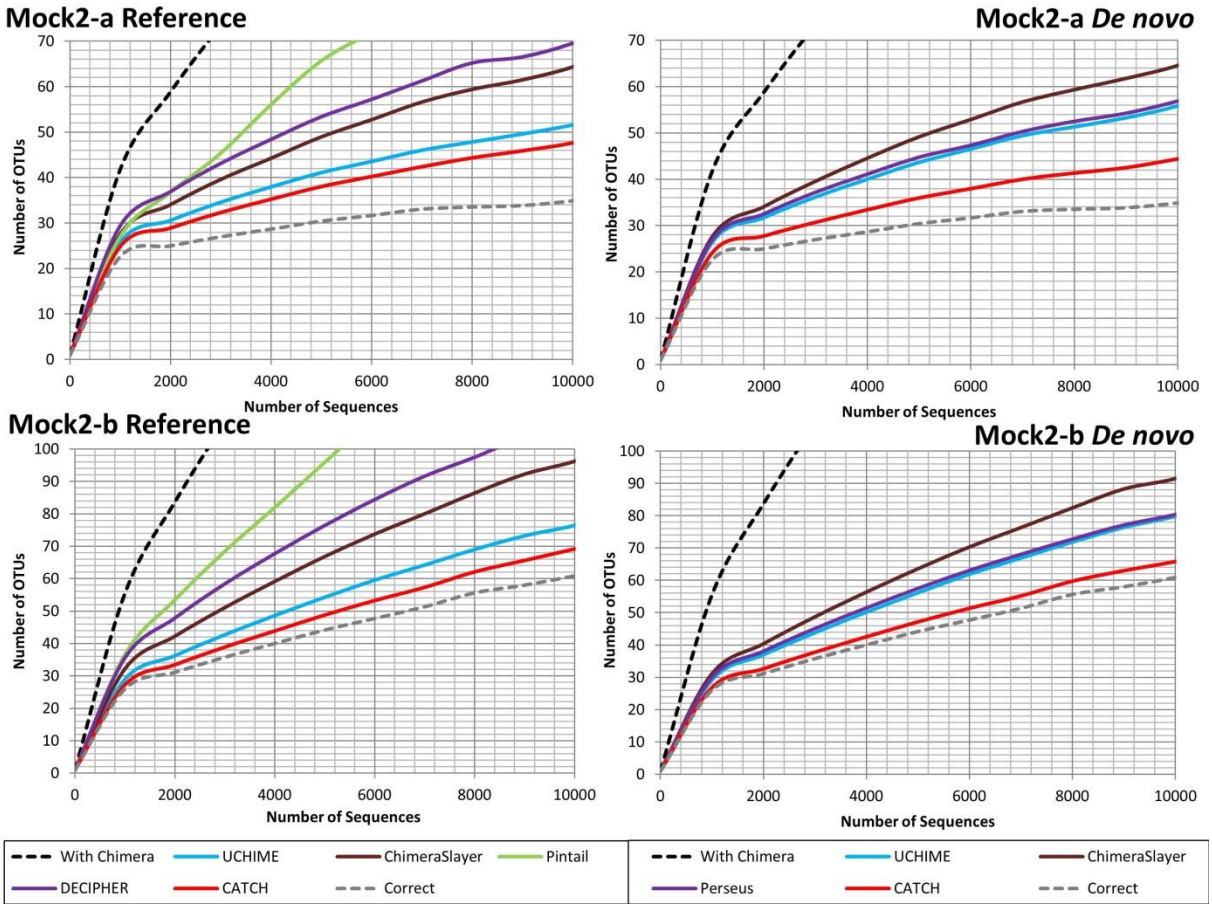


Figure S7: *Rarefaction curves (up to 10,000 reads) for Mock2 (a and b) including the results i) without any chimera removal ('with chimera'), ii) removing all chimeric sequences ('correct'), and iii) with chimera removal using all tested chimera detection tools (reference-based approaches on the left side and de novo approaches on the right side), averaged over all 69 samples.*

***Table S1**: Comparison of the different chimera detection algorithms.*

Tools	**UCHIME**	**Chimera Slayer**	**Pintail**	**DECIPHER**	**Perseus**
Available methods	Reference and *de novo*	Reference and *de novo*	Reference	Reference	*de novo*
Method	Reads divided into four non-over lapping segments, followed by three way alignment and scoring.	Each end used as seed to find closest parents. Followed by alignment and score calculation where cut-off is applied.	Calculating the differences between the query and the closest reference sequence. Next, expected differences between both (per position) are calculated.	The query sequence is classified to a reference phylogenetic group. A sliding window is used to split the query into different fragments, then depending on their presence inside/outside the assigned group, the query can be classified as chimeric or non-chimeric.	Uses pairwise alignment to identify the possible parents and breakpoints, for the other reads with abundance at least equal to the query. Performs three way alignment followed by scoring.
Look up database	Reference: Gold database. *De novo*: Equal or more abundant reads	Reference: Gold database. *De novo*: Equal or more abundant reads	Reference: Gold database.	RDP release 10, update 22 (+/- silva and greengenes).	Equal or more abundant reads

***Table S2**: Overview of the performance of the different regression kernels trained and tested in WEKA according to the procedure described in Material and Methods section. The upper table gives the sensitivity and specificity values obtained on the testing data for the reference based implementation, the lower table gives the overview for the* de novo *implementation. The selected kernel is highlighted.*

Reference	Linear Regression	Multi-Layer Perceptron	SVM linear	SVM PUK	Decision Stump	Random Forest
Senstivity	0.54	0.80	0.80	0.85	0.59	0.83
Specificity	**0.90**	**0.90**	**0.90**	**0.90**	**0.75**	**0.85**
MCC	0.41	0.66	0.66	0.71	0.31	0.64
Accuracy	0.65	0.83	0.83	0.86	0.64	0.83

De novo	Linear Regression	Multi-Layer Perceptron	SVM linear	SVM PUK	Decision Stump	Random Forest
Senstivity	0.89	0.91	0.91	0.92	0.80	0.89
Specificity	**0.93**	**0.93**	**0.93**	**0.93**	**0.94**	**0.93**
MCC	0.78	0.80	0.81	0.83	0.68	0.78
Accuracy	0.90	0.91	0.92	0.93	0.84	0.90

Table S3: *Illustration of the principle component analysis results, where all components explaining 95% of the variation in both the reference and* de novo *CATCh implementation are shown, together with the relative influence of each feature. For the reference based CATCh, seven components were needed (C1-C7) and three components (C1-C3) for the* de novo*. The last row illustrates the proportion of the variation explained by each principal component.*

	Reference							***De novo***			
	C1	**C2**	**C3**	**C4**	**C5**	**C6**	**C7**		**C1**	**C2**	**C3**
UCHIME Result	-0.02	+0.26	+0.75	-0.57	-0.11	+0.14	+0.06	**UCHIME Result**	+0.42	+0.17	-0.33
UCHIME Score	+ .36	+ .42	+ .07	+ .05	-0.02	-0.77	-0.28	**UCHIME Score**	+0.35	-0.85	-0.36
ChimeraSlayer Results	+ .46	+ .29	-0.06	+ .20	-0.30	+ .13	+ .74	**ChimeraSlayer Results**	+0.41	-0.10	+0.52
ChimeraSlayer Score	+ .51	+ .06	-0.06	+ .07	-0.40	+ .47	-0.57	**ChimeraSlayer Score**	+0.40	-0.07	+0.62
Pintail Score	-0.41	+ .46	+ .04	+ .31	-0.14	+ .00	+ .05	**Perseus Result**	+0.43	+ .35	-0.24
Pintail Std.	-0.40	+0.46	-0.01	+0.27	-0.15	+0.22	-0.21	**Perseus Score**	+0.43	+0.33	-0.23
Pintail Result	-0.14	+0.25	-0.63	-0.67	-0.25	-0.04	+0.04				
DECIPHER	+0.23	+0.43	-0.14	-0.08	+0.80	+0.31	-0.02				
Expressed Variation	0.31	0.23	0.13	0.11	0.09	0.06	0.05	**Expressed Variation**	0.81	0.08	0.06
Cumulative variation	0.31	0.54	0.67	0.78	0.87	0.93	0.98		0.81	0.89	0.95

Table S4*: Illustration of several classifiers built using all tools except one (either reference based tools or* de novo *based tools), using the same training dataset, and tested on Mock2-a. The name of the algorithm mentioned in the first row, is the algorithm which is omitted in creating the corresponding classifier. After adjusting the cut-off to give the same specificity as CATCh, we observed that CATCh (reference and* de novo*) integrating all available algorithms, provided the highest sensitivity.*

	Reference					***de novo***			
	Original	**UCHIME**	**ChimeraSlayer**	**Pintail**	**DECIPHER**	**Original**	**UCHIME**	**ChimeraSlayer**	**Perseus**
Sensitivity	0.82	0.76	0.80	0.69	0.38	0.85	0.82	0.82	0.79
Specificity	0.92	0.92	0.92	0.92	0.92	0.91	0.91	0.91	0.91
MCC	0.70	0.63	0.67	0.55	0.32	0.72	0.69	0.68	0.66
Accuracy	0.86	0.81	0.84	0.76	0.55	0.87	0.85	0.85	0.84

Table S5*: Comparison between different individual tools, the union of all results, and CATCh for A) reference-based algorithms and B)* de novo *algorithms. The reference-based implementations were tested on 25% of the Titanium dataset (this dataset is not used for training the reference-based CATCh classifier). For the* de novo *implementations, testing was performed on the F01QS4Z01_rep2_v13 subset from Mock2-b.*

A) Reference-based

	UCHIME	ChimeraSlayer	Pintail	DECIPHER	Union	CATCh
Sensitivity	0.41	0.24	0.00	0.80	0.87	0.85
Specificity	0.90	0.95	1.00	0.86	0.81	0.90
MCC	0.32	0.23	-	0.63	0.66	0.72
Accuracy	0.57	0.46	0.31	0.82	0.85	0.87

B) *De novo*

	UCHIME	ChimeraSlayer	Perseus	Union	CATCh
Sensitivity	0.79	0.71	0.79	0.93	0.92
Specificity	0.93	0.91	0.94	0.87	0.93
MCC	0.66	0.56	0.67	0.79	0.82
Accuracy	0.83	0.77	0.83	0.91	0.92

Table S6*: Illustration of the performance of three trained classifiers: CATCh reference, CATCh* de novo *and CATCh merged, using the MOCK2-a and MOCK2-b datasets. The results show marginal improvement in sensitivity when using CATCh merged compared CATCH* de novo *however with the same cost in specificity in MOCK2-b and no improvement in MOCK2-a.*

	Mock2-a			**Mock2-b**		
	CATCh Reference	**CATCh *De novo***	**CATCh merged**	**CATCh Reference**	**CATCh *De novo***	**CATCh merged**
Sensitivity	0.82	0.85	0.85	0.93	0.95	0.96
Specificity	0.92	0.91	0.91	0.92	0.91	0.90
MCC	0.70	0.72	0.71	0.82	0.84	0.87
Accuracy	0.86	0.87	0.87	0.92	0.94	0.95

Table S7: *Performance of different reference-based tools when dealing with deletions, insertions, indels and mismatches for Simu4 data, and indels and substitutions for Simu2 data.*

		UCHIME	***ChimeraSlayer***	***Pintail***	***DECIPHER***	***CATCh***		***UCHIME***	***ChimeraSlayer***	***Pintail***	***DECIPHER***	***CATCh***
Simu4	***Del(1%)***	0.94	0.89	0.12	0.33	0.95	***Del(5%)***	0.75	0.81	0.04	0.16	0.90
	Ind(1%)	0.94	0.89	0.15	0.34	0.95	***Ind(5%)***	0.74	0.80	0.14	0.19	0.89
	Ins(1%)	0.95	0.89	0.18	0.34	0.95	***Ins(5%)***	0.85	0.79	0.33	0.18	0.91
	Mis(1%)	0.94	0.86	0.13	0.32	0.95	***Mis(5%)***	0.85	0.67	0.06	0.25	0.86
Simu2	***Ind(1%)***	0.63	0.50	0.21	0.42	0.84	***Ind(5%)***	0.36	0.26	0.26	0.16	0.67
	Sub(1%)	0.66	0.39	0.20	0.48	0.82	***Sub(5%)***	0.47	0.10	0.20	0.30	0.68

Table S8: Sensitivity, specificity, Mathew correlation coefficient (MCC) and accuracy values of all tested chimera prediction tools . The first three datasets are tested with both reference and de novo based tools. The second part of the table contains the datasets either analysed using reference or de novo tools. The last three rows give respectively the average sensitivity and specificity, and the average increase or decrease in sensitivity / specificity when compared with CATCh.

		Reference tools					***De novo* tools**				
Evaluation	**Datasets**	**UCHIME**	**ChimeraSlayer**	**Pintail**	**DECIPHER**	**CATCh**	**Datasets**	**UCHIME**	**ChimeraSlayer**	**Perseus**	**CATCh**
Sensitvity	Mock2-A	0.79	0.66	0.58	0.61	0.82	Mock2-A	0.74	0.66	0.76	0.85
Specificity		0.94	0.95	0.42	0.93	0.92		0.94	0.94	0.92	0.91
MCC		0.68	0.56	-0.01	0.49	**0.7**		0.63	0.55	0.62	**0.72**
Accuracy		0.84	0.76	0.52	0.72	**0.86**		0.81	0.76	0.81	**0.87**
Sensitvity	Mock2-B	0.90	0.77	0.58	0.70	0.93	Mock2-B	0.87	0.80	0.87	0.95
Specificity		0.94	0.96	0.46	0.93	0.92		0.94	0.94	0.93	0.91
MCC		0.79	0.66	0.04	0.56	**0.82**		0.76	0.67	0.75	**0.84**
Accuracy		0.91	0.82	0.54	0.76	**0.92**		0.89	0.83	0.89	**0.94**
Sensitivity	Mock3	0.75	0.63	0.06	0.61	0.81	Mock3	0.64	0.65	0.73	0.81
Specificity		0.95	0.99	1.00	0.98	0.94		0.98	0.97	0.98	0.96
MCC		0.65	0.63	NA	0.6	**0.69**		0.63	0.62	0.69	**0.74**
Accuracy		0.82	0.78	0.5	0.78	**0.84**		0.8	0.79	0.84	**0.87**
Sensitivitv	Simu1	0.56	0.51	0.15	0.45	0.70	Mock1	0.92	0.71	0.94	0.97
Specificity		1.00	1.00	0.88	1.00	0.96		1.00	1.00	1.00	1.00
MCC		0.64	0.6	0.04	0.55	**0.69**		0.7	0.42	0.76	0.87
Accuracy		0.8	0.77	0.54	0.75	**0.84**		0.93	0.74	0.94	0.98
Sensitivity	Simu2	0.69	0.53	0.21	0.48	0.81	Mock4	0.37	0.27	0.27	0.47
Specificity		1.00	0.99	0.80	0.99	0.98		0.96	0.98	0.94	0.91
MCC		1	0.99	0.8	0.99	**0.98**		0.33	0.28	0.22	0.35
Accuracy		0.4	0.29	0.01	0.27	**0.51**		0.54	0.47	0.45	0.59
Sensitivity	Simu3	0.83	0.66	0.27	0.87	0.94	Mock5	0.08	0.11	0.14	0.16
Specificity		0.99	1.00	0.86	0.98	0.97		1.00	0.95	0.99	0.98
MCC		0.86	0.76	0.15	0.88	**0.90**		0.24	0.08	0.27	0.26
Accuracy		0.95	0.91	0.71	0.95	**0.96**		0.85	0.81	0.85	0.85
Sensitivity	Simu4	0.95	0.90	0.17	0.27	0.95					
Specificity		1.00	0.99	0.86	1.00	0.97					
MCC		0.96	0.9	0.04	0.44	**0.91**					
Accuracy		0.98	0.96	0.62	0.75	**0.96**					
Sensitivitv	**Average**	**0.78**	**0.67**	**0.29**	**0.57**	**0.85**	**Average**	**0.60**	**0.53**	**0.62**	**0.70**
Specificity		**0.97**	**0.98**	**0.75**	**0.97**	**0.96**		**0.97**	**0.96**	**0.96**	**0.95**
Difference		**7 / -2**	**18 / -2**	**56 / 21**	**28 / -1**	**-**		**10 / -2**	**17 / -1**	**8 / -1**	**-**
MCC		0.71	0.63	0.05	0.54	0.75		0.55	0.44	0.55	0.63
Accuracy		0.86	0.80	0.53	0.75	0.89		0.80	0.73	0.80	0.85