Supplementary File 2 - Supplementary Results

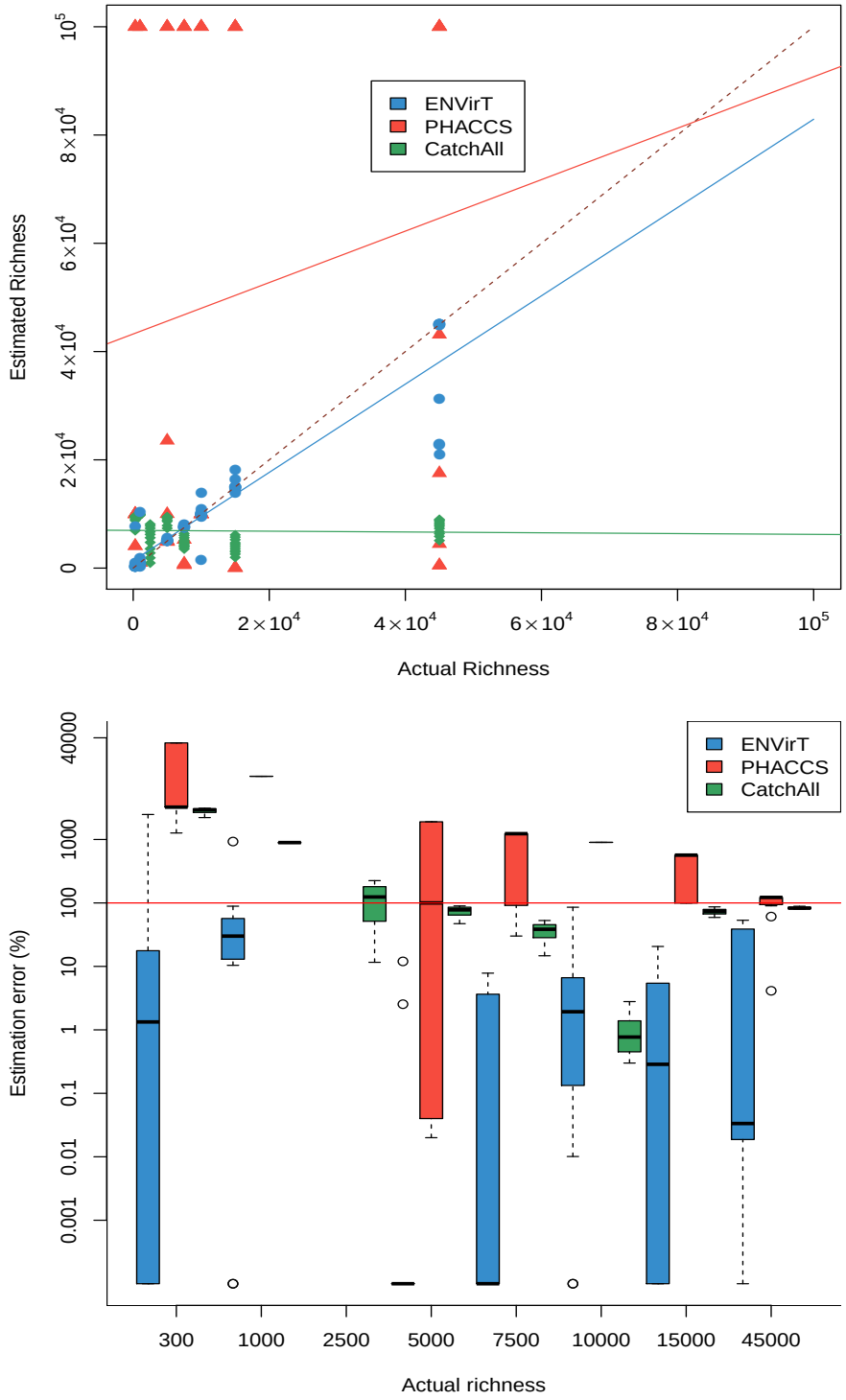

Figure S4: (top) Estimated richness (M) vs. True richness (M_0) under different average genome lengths (L); (bottom) Estimation error vs true richness. Note that estimation error values are plotted in a log2 scale for comparison between the larger errors produced by PHACCS in relation to ENVirT. On average we see that ENVirT performs up to 585% more accurately than PHACCS.

Table S1: CV(RMSE) values for L and M estimates for M = 300 and M = 10000 for different v values. The results are obtained by running the same dataset on ENVirT, ENVirT-FL and PHACCS. N/A - Not Applicable

		CV(RMSE) of L estimate			CV(RMSE) of M estimate		
Scenario	-log(v)	ENVirT	ENVirT-FL	PHACCS	ENVirT	ENVirT-FL	PHACCS
M = 300	4	0.00002	N/A	N/A	0.00000	0.00000	0.00000
	3.3	0.00007	N/A	N/A	0.00000	0.00000	0.00000
	3	0.00935	N/A	N/A	0.00913	0.00000	0.00000
	2.3	0.01828	N/A	N/A	0.01814	0.00000	0.00365
	2	0.02894	N/A	N/A	0.02896	0.00258	0.00548
	1.3	0.13252	N/A	N/A	0.14370	0.00882	0.02098
	1	0.28470	N/A	N/A	0.24046	0.02961	0.06672
M=10000	4	0.00018	N/A	N/A	0.00024	0.00000	0.00023
	3.3	0.00105	N/A	N/A	0.00130	0.00021	0.00112
	3	0.01595	N/A	N/A	0.00252	0.00042	0.00266
	2.3	0.03104	N/A	N/A	0.01363	0.01251	0.01246
	2	0.03581	N/A	N/A	0.01766	0.01468	0.01524
	1.3	0.09553	N/A	N/A	0.12784	0.11198	0.13040
	1	0.20204	N/A	N/A	0.20084	0.20102	0.22287

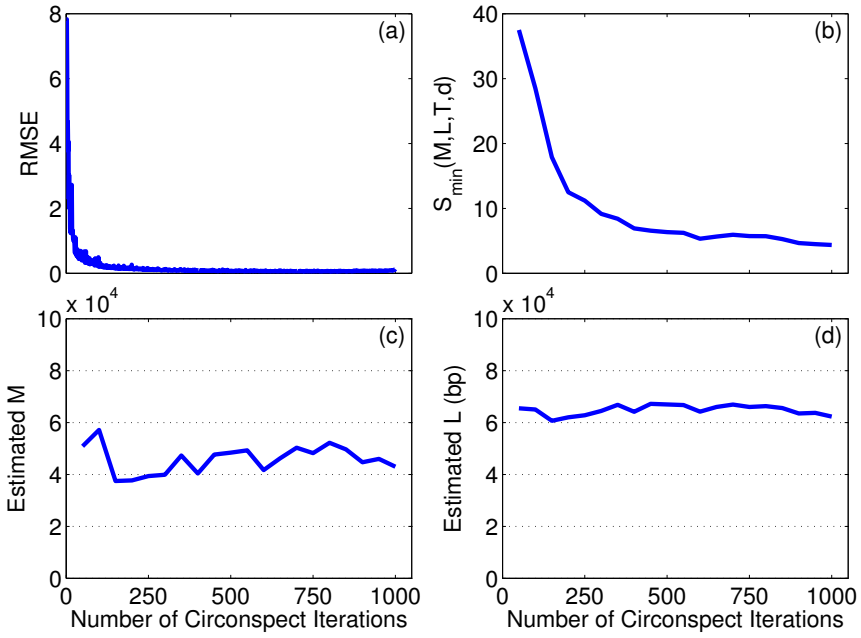

Figure S5: (a) Root Mean Squared Error (RMSE) between the average contig two consecutive iterations, (b) S(M,L,T,d) corresponding to the estimates, (c) Estimated M and (d) Estimated L (bp): produced by ENVirT under the contig spectrum of Lake Bourget averaged over different numbers of Circonspect iterations.

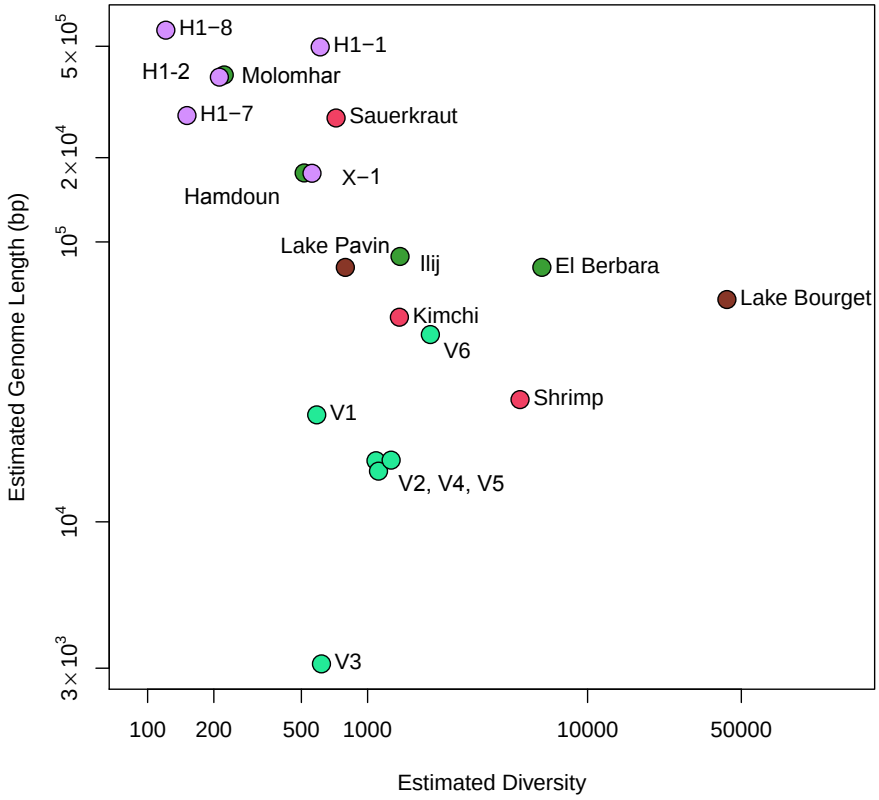

Figure S6: Estimated richness and average genome length as generated by ENVirT for 20 experimental metaviromes.

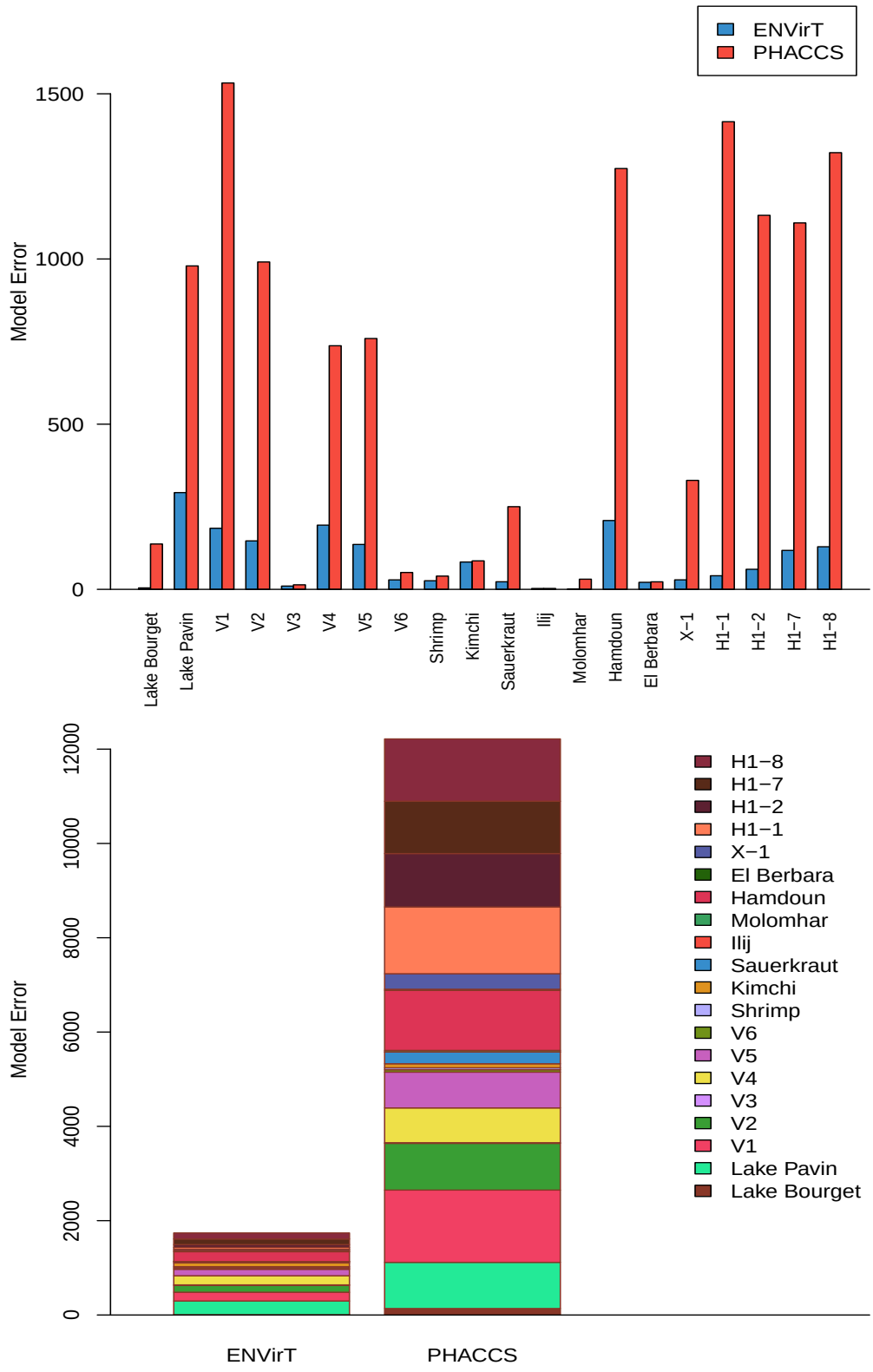

Figure S7: (top) Comparison of the residual model error as given by S_{min} between ENVirT and PHACCS+GAAS/BLAST; (bottom) The cumulative model error S_{min} for all 20 experimental metaviromes anlaysed by both ENVirT and PHACCS+GAAS/BLAST.

Table S2: Performance of ENVirT in comparison to standard GA algorithm on simulated contig spectra.

Input parameters (expected result)						Estimated values by ENVirT					Estimated values by GA without niching				
L_0	M_0	T_0	d_0	Evenness	f_{max}	L	M	T	d	S_{min}	L	M	T	d	S_{min}
12500	300	exp	0.030	0.790	2.956%	12500	300	exp	0.030	0.00x10 ^{0}	39500	12400	exp	0.095	3.49x10^{-2}
12500	1000	log	0.900	0.995	0.661%	14972	838	log	0.893	6.56x10^{-3}	310000	100	lgn	1.063	2.59x10 ^{1}
12500	5000	lgn	2.500	0.655	11.849%	12500	5000	lgn	2.500	0.00x10 ^{0}	12500	5000	lgn	2.500	0.00x10 ^{0}
12500	10000	pl	0.700	0.913	1.997%	12500	10000	pl	0.700	0.00x10 ^{0}	29500	1400	log	1.911	6.38x10 ^{0}
50000	300	exp	0.030	0.790	2.956%	50000	300	exp	0.030	0.00x10 ^{0}	41000	100	pl	0.378	1.53x10 ^{1}
50000	1000	log	0.900	0.995	0.661%	50000	1000	log	0.900	0.00x10 ^{0}	100500	600	lgn	0.531	3.48x10^{-2}
50000	5000	lgn	2.500	0.655	11.849%	50000	5000	lgn	2.500	0.00x10 ^{0}	50000	5100	lgn	2.506	1.92x10^{-2}
50000	10000	pl	0.700	0.913	1.997%	52787	10175	pl	0.707	1.72x10^{-3}	41000	9800	pl	0.677	2.22x10^{-2}
125000	300	exp	0.030	0.790	2.956%	125000	300	exp	0.030	0.00x10 ^{0}	58500	11000	exp	0.014	2.70x10^{-2}
125000	1000	log	0.900	0.995	0.661%	125000	1000	log	0.900	0.00x10 ^{0}	69000	1800	log	0.943	3.94x10^{-4}
125000	5000	lgn	2.500	0.655	11.849%	125000	5000	lgn	2.500	0.00x10 ^{0}	125000	5000	lgn	2.500	0.00x10 ^{0}
125000	10000	pl	0.700	0.913	1.997%	116341	9824	pl	0.691	1.96x10^{-4}	203000	15000	lgn	1.922	9.34x10^{-1}
300000	300	exp	0.030	0.790	2.956%	300000	300	exp	0.030	0.00x10 ^{0}	67000	400	lgn	0.543	5.36x10^{-2}
300000	1000	log	0.900	0.995	0.661%	217303	1373	log	0.899	1.26x10^{-7}	156000	1900	log	0.931	1.93x10^{-5}
300000	5000	lgn	2.500	0.655	11.849%	300000	5000	lgn	2.500	0.00x10 ^{0}	310000	7400	lgn	2.635	1.09x10^{-1}
300000	10000	pl	0.700	0.913	1.997%	277000	9800	pl	0.690	3.00x10^{-5}	77000	5600	log	1.658	2.97x10^{-2}

Contig spectra were generated with parameters: $R = 10000$, $r = 100bp$ and $o = 35bp$. pl = power-law distribution, exp = exponential distribution, log = logarithmic distribution and lgn = lognormal distribution. f_{max} = relative abundance of the dominant genotype. S_{min} = the value of the cost function corresponding to the estimated values of M, L, T and d. GA = Genetic Algorithm. We chose $M_{LB} = 1, M_{UB} = 15000, L_{LB} = 10000, L_{UB} = 310000, d_{LB} = 0.01$ and $d_{UB} = 5$ for both ENVirT and GA without niching. In order to apply the second niching strategy of ENVirT, we chose $N_L = 29$.