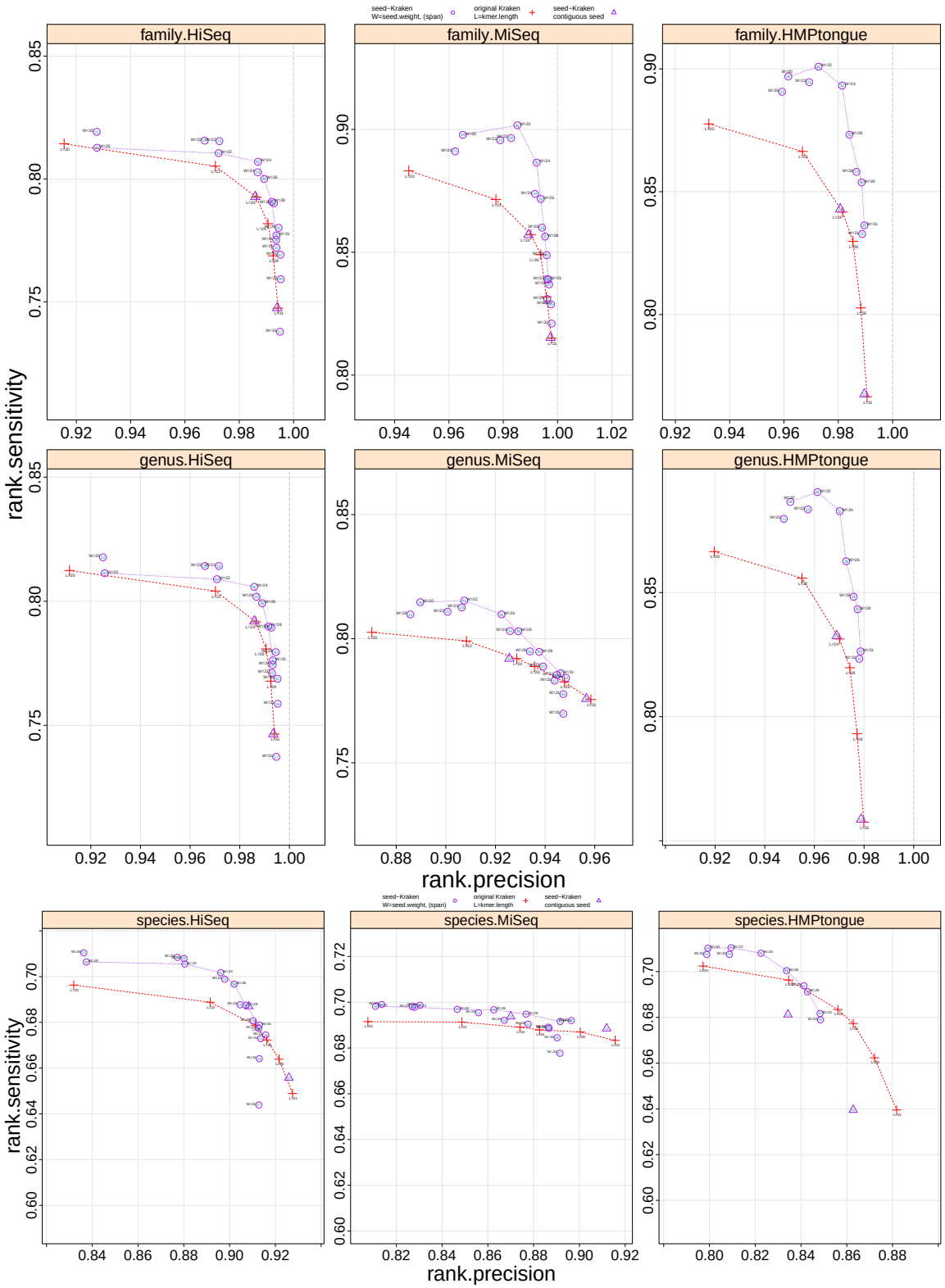

Figure 1: Performance of SEED-KRAKEN vs original KRAKEN with various seeds of different weights and spans. Sensitivity and precision obtained from experiments on HiSeq, MiSeq and HMPtongue datasets measured at 3 taxonomic rank levels: family, genus, species.

On Figure 1 we present results (receiver operating characteristic - ROC) from read classification experiments, which we conducted on 3 data sets HiSeq, MiSeq, HMPtongue, and we measured classification sensitivity (correctly classified/all reads to be classified), and precision (correctly classified/all attempted classifications), at 3 taxonomic levels: family, genus, species. Color filled circles present our selection of (approximately) best performing seeds.

SEED-KRAKEN uses slightly different assignment algorithm than KRAKEN.[1] SEED-KRAKEN indexes both k-mer and its inverse, then each read is processed in direct and complementary directions separately, then the result which produced more hits is selected. The effect of the assignment algorithm modification can be observed by comparing contiguous seed SEED-KRAKEN runs with corresponding original KRAKEN runs. At the level of genus and family ranks precision drops very slightly. At the level of species rank the drop in precision is more evident, especially high on HMPtongue data, at the same time on MiSeq and HiSeq data sets an increase in sensitivity is present. This drop in precision results from more classifications being made at the level of species, which would otherwise be classified at a higher rank by the original algorithm. This naturally increases sensitivity on HiSeq, and MiSeq, however not on HMPtongue, supposedly because of a specific database against which we tested HMPtongue (915 bacterial genomes, 3.3 GB in size, plus a selected subset of HMREFG HMP Reference Genome Database consisting of 1602 sequences, 0.5GB in size).

Considering span of a seed, for seeds of smaller weights 20, 22 shorter spans (+3 to weight) perform a bit worse than longer span S=31, with the exception of HiSeq data set, where the opposite occurs. For weights 24 and up, extending a seed span above 31 generally results with a drop in sensitivity, but there are exceptions like seed W=31, S=35 performing with a better sensitivity than W=31,S=33. Possibly there exists an optimal span for each weight of a seed with respect to sensitivity, but our data is not sufficient to deduce it, and suggests that such an optimal span would be dependent on a data set. The drop in sensitivity resulting from span increase usually is associated with precision increase, but there is a law of diminishing returns at work with the heaviest seed we tested of W=31.

Generally we conclude that seeds can improve ROC curve characteristics at the level of classification of genus and family ranks, for HMPtongue, HiSeq, MiSeq, and simBA5(not charted) data sets, in the tested range of k-mer weights 20-31. At the species level the situation is complicated by the greater influence of the change in the assignment algorithm, although we can conclude that seeds allow for increased sensitivity, but the question weather it can be increased in the high precision regime remains open.

On simBA5 data set (not charted, see tables) the results on genus and family follow the tendency of other data sets. In our experiment this data set results with low sensitivity, as the database we used was reduced in half comparing to original Kraken database (915 bacterial genomes, 3.3 GB in size), while simBA5 consists of 10K reads from a wide range of types of ogranisms.

[1] This is because a complement of a spaced k-mer does not match the seed but its inverse.

	exp.name	seedspan	seedweight	seed.seq	sens.HiSeq	prec.HiSeq	sens.HMPtongue	prec.HMPtongue	sens.MiSeq	prec.MiSeq	sens.simBA5	prec.simBA5
1	orig-Kraken-l20	20	20	`####################`	0.6963	0.8319	0.7024	0.7972	0.6915	0.8076	0.4743	0.6564
2	seed-Kraken-w20-l23-hitkarel_short	23	20	`######-###-#####-######`	0.7104	0.8363	0.7075	0.7988	0.6990	0.8136	0.4822	0.6689
3	seed-Kraken-w20-l31-laurent	31	20	`#######--#-#---#-##---##-######`	0.7064	0.8374	0.7102	0.7993	0.6982	0.8109	0.4750	0.6791
4	orig-Kraken-l22	22	22	`######################`	0.6888	0.8915	0.6963	0.8346	0.6913	0.8487	0.4719	0.7615
5	seed-Kraken-w22-l25-hitkarel_short	25	22	`######-####-#####-#######`	0.7084	0.8772	0.7075	0.8087	0.6981	0.8270	0.4827	0.7388
6	seed-Kraken-w22-l28-hitkarel_short	28	22	`#######--##-###-#-####-#####`	0.7079	0.8801			0.6987	0.8304	0.4796	0.7393
7	seed-Kraken-w22-l31-laurent	31	22	`######-###--#-##-#--##--#######`	0.7055	0.8804	0.7104	0.8094	0.6978	0.8280	0.4768	0.7392
8	orig-Kraken-l24	24	24	`########################`	0.6788	0.9113	0.6834	0.8561	0.6891	0.8740	0.4677	0.8281
9	seed-Kraken-w24-l24	24	24	`########################`	0.6870	0.9084	0.6812	0.8343	0.6938	0.8700	0.4738	0.8115
10	seed-Kraken-w24-l31-laurent	31	24	`#######-##-##--##-#-#-#########`	0.7017	0.8961	0.7080	0.8225	0.6969	0.8466	0.4739	0.7914
11	seed-Kraken-w24-l34-laurent	34	24	`########---##-#-##-#---###-#######`	0.6988	0.8979			0.6954	0.8559	0.4702	0.8057
12	orig-Kraken-l26	26	26	`##########################`	0.6723	0.9163	0.6773	0.8628	0.6879	0.8826	0.4606	0.8507
13	seed-Kraken-w26-l29-hitkarel_short_mac	29	26	`########-####-######-########`	0.6967	0.9020	0.7004	0.8337	0.6967	0.8626	0.4746	0.8157
14	seed-Kraken-w26-l31-laurent	31	26	`##########-#-##-###--##########`	0.6924	0.9045	0.6971	0.8353	0.6954	0.8682	0.4707	0.8256
15	seed-Kraken-w26-l38-laurent	38	26	`#####-##-#-#-##-###-#-#---###--#######`	0.6877	0.9046			0.6921	0.8672	0.4534	0.8271
16	orig-Kraken-l28	28	28	`############################`	0.6639	0.9216	0.6623	0.8721	0.6870	0.9004	0.4505	0.8704
17	seed-Kraken-w28-l31-hitkarel_short	31	28	`#######-######-#####-##########`	0.6875	0.9071	0.6938	0.8413	0.6948	0.8767	0.4677	0.8318
18	seed-Kraken-w28-l40-laurent	40	28	`######-##-#-#-##-###-#-##---###--#######`	0.6807	0.9103	0.6912	0.8428	0.6904	0.8776	0.4420	0.8381
19	orig-Kraken-l31	31	31	`##############################`	0.6489	0.9275	0.6395	0.8818	0.6833	0.9156	0.4317	0.8908
20	seed-Kraken-w31-l31	31	31	`##############################`	0.6558	0.9259	0.6395	0.8628	0.6885	0.9119	0.4372	0.8728
21	seed-Kraken-w31-l33-hitkarel31	33	31	`##########-#########-#########`	0.6745	0.9158			0.6920	0.8965	0.4560	0.8551
22	seed-Kraken-w31-l35-hitkarel31	35	31	`##########-####-######-#####-#######`	0.6788	0.9130	0.6789	0.8485	0.6916	0.8916	0.4551	0.8552
23	seed-Kraken-w31-l38-hitkarel31	38	31	`#########-####-#--##-#####-####-######`	0.6774	0.9127			0.6891	0.8865	0.4465	0.8462
24	seed-Kraken-w31-l42-hitkarel	42	31	`#######-###-#-###-####-#--###-#--##-######`	0.6729	0.9136	0.6817	0.8483	0.6886	0.8867	0.4265	0.8515
25	seed-Kraken-w31-l46-laurent	46	31	`###-###-#-##-#-##-##-###-##-##---###--###-####`	0.6641	0.9130			0.6845	0.8903	0.4051	0.8578
26	seed-Kraken-w31-l57-qr	57	31	`##-###---#-#-#---###-##-----##--#-##-#-######-#-##-#--##`	0.6439	0.9128			0.6777	0.8915	0.3386	0.8578

Table 1: Experiments results at the 'species' level.

	exp.name	seedspan	seedweight	seed.seq	sens.HiSeq	prec.HiSeq	sens.HMPtongue	prec.HMPtongue	sens.MiSeq	prec.MiSeq	sens.simBA5	prec.simBA5
1	orig-Kraken-l20	20	20	`####################`	0.8124	0.9114	0.8665	0.9197	0.8026	0.8702	0.6130	0.7761
2	seed-Kraken-w20-l23-hitkarel_short	23	20	`######-###-#####-######`	0.8177	0.9249	0.8797	0.9477	0.8098	0.8857	0.6232	0.8093
3	seed-Kraken-w20-l31-laurent	31	20	`#######--#-#---#-##---##-######`	0.8113	0.9256	0.8865	0.9503	0.8147	0.8897	0.6119	0.8153
4	orig-Kraken-l22	22	22	`######################`	0.8041	0.9702	0.8558	0.9550	0.7991	0.9083	0.6048	0.8814
5	seed-Kraken-w22-l25-hitkarel_short	25	22	`######-####-#####-#######`	0.8142	0.9660	0.8835	0.9574	0.8109	0.9007	0.6204	0.8860
6	seed-Kraken-w22-l28-hitkarel_short	28	22	`#######--##-###-#-####-#####`	0.8142	0.9716			0.8126	0.9064	0.6184	0.8886
7	seed-Kraken-w22-l31-laurent	31	22	`######-###--#-##-#--##--#######`	0.8089	0.9708	0.8905	0.9612	0.8154	0.9074	0.6158	0.8861
8	orig-Kraken-l24	24	24	`########################`	0.7917	0.9865	0.8313	0.9701	0.7920	0.9286	0.5905	0.9407
9	seed-Kraken-w24-l24	24	24	`########################`	0.7920	0.9859	0.8325	0.9690	0.7920	0.9256	0.5907	0.9368
10	seed-Kraken-w24-l31-laurent	31	24	`#######-##-##--##-#-#-#########`	0.8058	0.9858	0.8828	0.9702	0.8098	0.9224	0.6018	0.9315
11	seed-Kraken-w24-l34-laurent	34	24	`########---##-#-##-#---###-#######`	0.8018	0.9867			0.8031	0.9259	0.5953	0.9434
12	orig-Kraken-l26	26	26	`##########################`	0.7808	0.9906	0.8197	0.9743	0.7888	0.9357	0.5766	0.9559
13	seed-Kraken-w26-l29-hitkarel_short_mac	29	26	`########-####-######-########`	0.7992	0.9890	0.8626	0.9728	0.8030	0.9292	0.5976	0.9513
14	seed-Kraken-w26-l31-laurent	31	26	`##########-#-##-###--##########`	0.7949	0.9910	0.8543	0.9741	0.7977	0.9316	0.5882	0.9539
15	seed-Kraken-w26-l38-laurent	38	26	`#####-##-#-#-##-###-#-#---###--#######`	0.7894	0.9927			0.7949	0.9339	0.5704	0.9613
16	orig-Kraken-l28	28	28	`############################`	0.7678	0.9925	0.7931	0.9772	0.7826	0.9478	0.5587	0.9661
17	seed-Kraken-w28-l31-hitkarel_short	31	28	`#######-######-#####-##########`	0.7899	0.9915	0.8483	0.9758	0.7947	0.9376	0.5835	0.9597
18	seed-Kraken-w28-l40-laurent	40	28	`######-##-#-#-##-###-#-##---###--#######`	0.7795	0.9945	0.8433	0.9774	0.7888	0.9392	0.5519	0.9665
19	orig-Kraken-l31	31	31	`##############################`	0.7465	0.9940	0.7575	0.9799	0.7755	0.9584	0.5297	0.9747
20	seed-Kraken-w31-l31	31	31	`##############################`	0.7465	0.9935	0.7586	0.9788	0.7758	0.9566	0.5301	0.9734
21	seed-Kraken-w31-l33-hitkarel31	33	31	`##########-#########-#############`	0.7711	0.9930			0.7842	0.9484	0.5615	0.9698
22	seed-Kraken-w31-l35-hitkarel31	35	31	`##########-####-######-#####-#######`	0.7762	0.9933	0.8233	0.9781	0.7861	0.9462	0.5615	0.9716
23	seed-Kraken-w31-l38-hitkarel31	38	31	`#########-####-#--##-#####-####-######`	0.7744	0.9933			0.7854	0.9447	0.5539	0.9707
24	seed-Kraken-w31-l42-hitkarel	42	31	`#######-###-#-###-####-#--###-#--##-######`	0.7688	0.9952	0.8263	0.9786	0.7831	0.9438	0.5277	0.9719
25	seed-Kraken-w31-l46-laurent	46	31	`###-###-#-##-#-##-##-###-##-##---###--###-####`	0.7587	0.9953			0.7777	0.9473	0.4999	0.9753
26	seed-Kraken-w31-l57-qr	57	31	`##-###---#-#-#---###-##-----##--#-##-#-######-#-##-#--##`	0.7373	0.9947			0.7698	0.9473	0.4164	0.9753

Table 2: Experiments results at the 'genus' level.

	exp.name	seedspan	seedweight	seed.seq	sens.HiSeq	prec.HiSeq	sens.HMPtongue	prec.HMPtongue	sens.MiSeq	prec.MiSeq	sens.simBA5	prec.simBA5
1	orig-Kraken-l20	20	20	`####################`	0.8143	0.9155	0.8776	0.9323	0.8832	0.9452	0.6487	0.8139
2	seed-Kraken-w20-l23-hitkarel_short	23	20	`######-###-#####-######`	0.8192	0.9275	0.8907	0.9593	0.8911	0.9623	0.6577	0.8455
3	seed-Kraken-w20-l31-laurent	31	20	`#######--#-#---#-##---##-######`	0.8127	0.9276	0.8969	0.9616	0.8978	0.9651	0.6457	0.8499
4	orig-Kraken-l22	22	22	`######################`	0.8052	0.9712	0.8664	0.9668	0.8715	0.9773	0.6363	0.9114
5	seed-Kraken-w22-l25-hitkarel_short	25	22	`######-####-#####-#######`	0.8156	0.9673	0.8946	0.9693	0.8956	0.9789	0.6535	0.9172
6	seed-Kraken-w22-l28-hitkarel_short	28	22	`#######--##-###-#-####-#####`	0.8154	0.9727			0.8965	0.9829	0.6518	0.9223
7	seed-Kraken-w22-l31-laurent	31	22	`######-###--#-##-#--##--#######`	0.8105	0.9724	0.9009	0.9727	0.9017	0.9852	0.6487	0.9170
8	orig-Kraken-l24	24	24	`########################`	0.7926	0.9863	0.8417	0.9816	0.8572	0.9902	0.6191	0.9651
9	seed-Kraken-w24-l24	24	24	`########################`	0.7927	0.9859	0.8429	0.9807	0.8571	0.9893	0.6192	0.9636
10	seed-Kraken-w24-l31-laurent	31	24	`#######-##-##--##-#-#-#########`	0.8070	0.9869	0.8932	0.9814	0.8865	0.9923	0.6338	0.9613
11	seed-Kraken-w24-l34-laurent	34	24	`########---##-#-##-#---###-#######`	0.8027	0.9868			0.8738	0.9917	0.6242	0.9703
12	orig-Kraken-l26	26	26	`##########################`	0.7818	0.9906	0.8298	0.9854	0.8490	0.9938	0.6026	0.9775
13	seed-Kraken-w26-l29-hitkarel_short_mac	29	26	`########-####-######-########`	0.8000	0.9892	0.8732	0.9841	0.8717	0.9938	0.6259	0.9769
14	seed-Kraken-w26-l31-laurent	31	26	`##########-#-##-###--##########`	0.7962	0.9919	0.8646	0.9853	0.8609	0.9931	0.6151	0.9784
15	seed-Kraken-w26-l38-laurent	38	26	`#####-##-#-#-##-###-#-#---###--#######`	0.7901	0.9928			0.8600	0.9943	0.5946	0.9814
16	orig-Kraken-l28	28	28	`############################`	0.7687	0.9925	0.8027	0.9883	0.8318	0.9959	0.5818	0.9845
17	seed-Kraken-w28-l31-hitkarel_short	31	28	`#######-######-#####-##########`	0.7907	0.9920	0.8581	0.9867	0.8564	0.9954	0.6097	0.9829
18	seed-Kraken-w28-l40-laurent	40	28	`######-##-#-#-##-###-#-##---###--#######`	0.7801	0.9945	0.8538	0.9886	0.8488	0.9960	0.5749	0.9870
19	orig-Kraken-l31	31	31	`##############################`	0.7474	0.9943	0.7665	0.9906	0.8150	0.9978	0.5497	0.9907
20	seed-Kraken-w31-l31	31	31	`##############################`	0.7475	0.9940	0.7676	0.9896	0.8152	0.9974	0.5500	0.9905
21	seed-Kraken-w31-l33-hitkarel31	33	31	`##########-#########-############`	0.7720	0.9937			0.8310	0.9962	0.5840	0.9895
22	seed-Kraken-w31-l35-hitkarel31	35	31	`##########-####-######-#####-#######`	0.7770	0.9937	0.8328	0.9888	0.8389	0.9966	0.5841	0.9902
23	seed-Kraken-w31-l38-hitkarel31	38	31	`#########-####-#--##-#####-####-######`	0.7752	0.9936			0.8391	0.9962	0.5765	0.9899
24	seed-Kraken-w31-l42-hitkarel	42	31	`#######-###-#-###-####-#--###-#--##-######`	0.7691	0.9952	0.8363	0.9896	0.8369	0.9969	0.5482	0.9909
25	seed-Kraken-w31-l46-laurent	46	31	`###-###-#-##-#-##-##-###-##-##---###--###-####`	0.7592	0.9953			0.8288	0.9976	0.5195	0.9930
26	seed-Kraken-w31-l57-qr	57	31	`##-###---#-#-#---###-##-----##--#-##-#-######-#-##-#--##`	0.7378	0.9950			0.8210	0.9978	0.4325	0.9925

Table 3: Experiments results at the 'family' level.

Performance of paired-end reads

	exp.name	seedspan	seedweight	seed.seq	sens.HMPtongue	prec.HMPtongue	sens.HMPtongue.paired	prec.HMPtongue.paired
2	seed-Kraken-w20-l23-hitkarel_short	23	20	`######-###-#####-######`	0.7075	0.7988	0.7154	0.8000
5	seed-Kraken-w22-l25-hitkarel_short	25	22	`######-####-#####-#######`	0.7075	0.8087	0.7154	0.8174
8	seed-Kraken-w24-l24	24	24	`########################`	0.6812	0.8343	0.6878	0.8442
11	seed-Kraken-w26-l29-hitkarel_short_mac	29	26	`########-####-######-########`	0.7004	0.8337	0.7064	0.8401
14	seed-Kraken-w28-l31-hitkarel_short	31	28	`#######-######-#####-##########`	0.6938	0.8413	0.6989	0.8480
17	seed-Kraken-w31-l31	31	31	`###############################`	0.6395	0.8628	0.6435	0.8701
18	seed-Kraken-w31-l35-hitkarel31	35	31	`#########-####-######-#####-#######`	0.6789	0.8485	0.6840	0.8542

Table 4: Experiments results at the 'species' level.

Performance of paired-end reads

	exp.name	seedspan	seedweight	seed.seq	sens.HMPtongue	prec.HMPtongue	sens.HMPtongue.paired	prec.HMPtongue.paired
2	seed-Kraken-w20-l23-hitkarel_short	23	20	`######-###-#####-######`	0.8797	0.9477	0.8818	0.9391
5	seed-Kraken-w22-l25-hitkarel_short	25	22	`######-####-#####-#######`	0.8835	0.9574	0.8841	0.9572
8	seed-Kraken-w24-l24	24	24	`########################`	0.8325	0.9690	0.8316	0.9690
11	seed-Kraken-w26-l29-hitkarel_short_mac	29	26	`########-####-######-########`	0.8626	0.9728	0.8624	0.9732
14	seed-Kraken-w28-l31-hitkarel_short	31	28	`#######-######-#####-##########`	0.8483	0.9758	0.8472	0.9759
17	seed-Kraken-w31-l31	31	31	`###############################`	0.7586	0.9788	0.7575	0.9790
18	seed-Kraken-w31-l35-hitkarel31	35	31	`#########-####-######-#####-#######`	0.8233	0.9781	0.8224	0.9784

Table 5: Experiments results at the 'genus' level.

Performance of paired-end reads

	exp.name	seedspan	seedweight	seed.seq	sens.HMPtongue	prec.HMPtongue	sens.HMPtongue.paired	prec.HMPtongue.paired
2	seed-Kraken-w20-l23-hitkarel_short	23	20	`######-###-#####-######`	0.8907	0.9593	0.8927	0.9513
5	seed-Kraken-w22-l25-hitkarel_short	25	22	`######-####-#####-#######`	0.8946	0.9693	0.8950	0.9693
8	seed-Kraken-w24-l24	24	24	`########################`	0.8429	0.9807	0.8419	0.9808
11	seed-Kraken-w26-l29-hitkarel_short_mac	29	26	`########-####-######-########`	0.8732	0.9841	0.8726	0.9844
14	seed-Kraken-w28-l31-hitkarel_short	31	28	`#######-######-#####-##########`	0.8581	0.9867	0.8570	0.9868
17	seed-Kraken-w31-l31	31	31	`###############################`	0.7676	0.9896	0.7665	0.9899
18	seed-Kraken-w31-l35-hitkarel31	35	31	`#########-####-######-#####-#######`	0.8328	0.9888	0.8320	0.9893

Table 6: Experiments results at the 'family' level.