l_i	Mean	Std. dev.
1	1.0193	0.1227
2	2.0006	0.1585
3	2.9934	0.2188
4	3.9962	0.3168
5	4.9550	0.3863
> 5	l_i	$0.03494 + l_i \cdot 0.06856$

Table 1. Empirical distribution of homopolymer lengths in 454 sequenced data from Balzer *et al.* (2010)

1 SIMULATION OF SPECIFIC 454 SEQUENCING ERROR

For simulating typical 454 sequencing errors we applied a similar approach as in grinder (`http://sourceforge.net/projects/biogrinder/`). The simulation works by scanning the QS for homopolymers (contiguous stretches of one or more occurances of the same character). Thereby the QS can be interpreted as a sequence of homoplymers: $c_1^{l_1} c_2^{l_2} ... c_n^{l_n}$, where $c_i^{l_i}$ denotes a homopolymer of character c_i and length $l_i > 0$. For example, the sequence CCGTTTACCAA corresponds to the homopolymer sequence $C^2 G^1 T^3 A^1 C^2 A^2$. The simulation generates a modified sequence $c_1^{m_1} c_2^{m_2} ... c_n^{m_n}$ where $m_i = rnd(l_i)$ is a normally distributed random value with mean values and standard deviations drawn from the empirical data in Table 1. Our implementation of the QS error generation method is available at `https://github.com/sim82/java_tools/blob/master/src/ml/SampleDistSubseq.java`.

2 MEMORY LAYOUT CONSIDERATIONS AND EXPERIMENTAL PARALLELIZATION

In the current implementation, the complete dynamic-programming matrix is kept in memory, whereas in theory only one row is actually needed for the sequential implementation. This —larger than necessary— memory footprint might have a negative impact on cache utilization/efficiency. An initial assessment using cachegrind (Nethercote and Seward, 2007) indicated that cache utilization is near-optimal, with an estimated L1 cache miss-rate of 0.6% (0.1% read, 3.6% write) for a subset of the 1604/200±60bp data-set and a L1 cache size of 32kB. For most data-sets used in our experiments the dynamic programming matrix entirely fits into the L2 or L3 caches of current CPUs.

We also developed a proof-of-concept parallelization using PThreads to exploit embarrasing parallelism by independently and simultaneously aligning QS against multiple ancestral sequences (insertion branches) in analogy to the coarse-grain parallel phase of the multi-grain EPA parallelization (Stamatakis *et al.*, 2010). Despite being in a proof-of-concept state, execution times of the parallelized part scale almost linearly up to 32 cores on a SUN x4600 multi-core system.

3 RAW RESULTS ON THE REMAINING DATA-SETS

		ND		EDN %	
D	MA	PaPaRa	HMM	PaPaRa	HMM
D150	ORIG	0.48	1.19 (2.49)	1.72	1.58 (0.92)
	MUSCLE	0.52	1.09 (2.09)	1.94	1.37 (0.71)
	MAFFT	0.50	1.19 (2.38)	1.67	1.40 (0.84)
	PRANK$_{+F}$	0.65	2.01 (3.09)	1.82	2.85 (1.56)
D218	ORIG	1.99	1.78 (0.90)	6.42	5.75 (0.89)
	MUSCLE	1.93	1.64 (0.85)	7.50	5.93 (0.79)
	MAFFT	1.81	1.90 (1.05)	6.05	6.07 (1.00)
	PRANK$_{+F}$	2.04	1.99 (0.97)	7.14	6.64 (0.93)
D500	ORIG	0.43	0.43 (0.99)	1.26	1.21 (0.96)
	MUSCLE	0.46	0.51 (1.11)	1.31	1.43 (1.09)
	MAFFT	0.47	0.49 (1.04)	1.33	1.36 (1.02)
	PRANK$_{+F}$	0.55	0.64 (1.17)	1.52	1.74 (1.14)
D628	ORIG	0.73	3.14 (4.32)	1.94	8.08 (4.16)
	MUSCLE	1.02	3.64 (3.57)	2.64	9.03 (3.42)
	MAFFT	0.42	2.66 (6.37)	1.00	4.68 (4.68)
	PRANK$_{+F}$	0.42	3.27 (7.87)	1.00	5.53 (5.53)
D714	ORIG	0.49	0.46 (0.95)	1.51	1.06 (0.70)
	MUSCLE	0.44	0.76 (1.73)	1.23	1.21 (0.98)
	MAFFT	0.40	0.66 (1.67)	1.41	1.31 (0.93)
	PRANK$_{+F}$	0.48	1.21 (2.53)	1.37	2.30 (1.68)
D855	ORIG	0.46	0.97 (2.11)	0.83	1.27 (1.53)
	MUSCLE	0.54	1.24 (2.27)	1.01	1.86 (1.84)
	MAFFT	0.51	0.62 (1.20)	0.87	0.90 (1.03)
	PRANK$_{+F}$	0.68	2.09 (3.07)	1.24	3.24 (2.60)
D1604	ORIG	0.23	1.25 (5.42)	0.63	1.19 (1.90)
	MUSCLE	0.38	1.26 (3.31)	0.80	1.36 (1.70)
	MAFFT	0.24	1.12 (4.60)	0.64	1.17 (1.84)
	PRANK$_{+F}$	0.38	2.34 (6.19)	0.90	2.28 (2.53)

Table 2. Placement accuracy for the two QS alignment methods on all data-sets using QS of length 100 ± 10 bp **without** simulated 454 errors.

4 EXPERIMENTS WITH SIMULATED DATA

For the sake of completeness, we also carried out additional experiments using simulated data. We used INDELible (v1.03; Fletcher and Yang, 2009) to generate/simulate realistic synthetic sequences based on the best-known ML tree for the largest —with respect to the number of taxa— real world data-set (D1604) in our study. The experiment/simulation was conducted as follows: Initially, we inferred the best-know ML tree using RAxML and the GTR+Γ model on the original MSA for data-set D1604. Then, we simulated a new MSA on this tree with sequence lengths of 1200bp with INDELible using the following parameter settings for the nucleotide substitution and indel models: the parameters of the GTR+Γ model as estimated by RAxML during the inference of the ML tree on the real-world MSA; the empirical base frequencies as estimated from the original —real-world— MSA; the indel model parameters as estimated using the `lambda.pl` script from Dawg (v1.2; Cartwright, 2005). The Power-Law Model (POW) for the indels was suggested by `lambda.pl` as being the most likely indel model. We chose the indel rate parameter, such that the simulated MSA contained $\approx$ 1600 alignment sites. Thus, the simulated MSA contained a relatively high amount of gaps compared to the corresponding real-world MSA, which should yield the QS alignment process harder:

D	MA	ND		EDN %	
		PaPaRa	HMM	PaPaRa	HMM
D150	ORIG	0.52	1.31 (2.52)	1.74	1.68 (0.96)
	MUSCLE	0.57	1.18 (2.07)	2.00	1.51 (0.75)
	MAFFT	0.58	1.29 (2.24)	1.69	1.49 (0.88)
	PRANK$_{+F}$	0.70	2.10 (3.02)	1.88	2.89 (1.54)
D218	ORIG	2.02	1.78 (0.88)	6.57	5.80 (0.88)
	MUSCLE	1.95	1.73 (0.88)	7.63	6.25 (0.82)
	MAFFT	1.86	1.92 (1.03)	6.21	6.14 (0.99)
	PRANK$_{+F}$	2.04	2.03 (0.99)	7.18	6.86 (0.96)
D500	ORIG	0.57	0.59 (1.03)	1.64	1.65 (1.01)
	MUSCLE	0.60	0.68 (1.13)	1.71	1.91 (1.12)
	MAFFT	0.62	0.69 (1.12)	1.74	1.91 (1.10)
	PRANK$_{+F}$	0.68	0.81 (1.20)	1.88	2.22 (1.18)
D628	ORIG	0.80	1.90 (2.39)	2.07	3.89 (1.88)
	MUSCLE	1.09	3.75 (3.44)	2.81	9.38 (3.34)
	MAFFT	0.47	2.78 (5.90)	1.11	5.00 (4.51)
	PRANK$_{+F}$	0.50	3.32 (6.68)	1.14	5.61 (4.92)
D714	ORIG	0.55	0.54 (0.99)	1.71	1.28 (0.75)
	MUSCLE	0.50	0.86 (1.70)	1.40	1.40 (1.00)
	MAFFT	0.45	0.75 (1.65)	1.58	1.55 (0.98)
	PRANK$_{+F}$	0.51	1.28 (2.50)	1.47	2.48 (1.68)
D855	ORIG	0.59	1.32 (2.24)	1.03	1.67 (1.62)
	MUSCLE	0.67	1.54 (2.29)	1.22	2.32 (1.90)
	MAFFT	0.66	1.03 (1.56)	1.11	1.50 (1.35)
	PRANK$_{+F}$	0.80	2.28 (2.85)	1.47	3.57 (2.43)
D1604	ORIG	0.28	1.35 (4.87)	0.71	1.35 (1.90)
	MUSCLE	0.43	1.35 (3.11)	0.87	1.48 (1.70)
	MAFFT	0.29	1.21 (4.12)	0.72	1.29 (1.80)
	PRANK$_{+F}$	0.41	2.43 (5.88)	0.95	2.41 (2.52)

Table 3. Placement accuracy for the two QS alignment methods on all datasets using QS of length 100 ± 10 bp **with** simulated 454 errors.

D	MA	ND		EDN %	
		PaPaRa	HMM	PaPaRa	HMM
D150	ORIG	0.46	0.62 (1.34)	1.26	1.00 (0.79)
	MUSCLE	0.46	0.52 (1.12)	1.44	0.64 (0.44)
	MAFFT	0.46	0.60 (1.29)	1.22	0.79 (0.65)
	PRANK$_{+F}$	0.52	1.08 (2.08)	1.53	1.49 (0.97)
D218	ORIG	1.57	1.06 (0.68)	5.53	3.61 (0.65)
	MUSCLE	1.69	0.98 (0.58)	6.94	4.08 (0.59)
	MAFFT	1.59	1.24 (0.78)	5.63	4.37 (0.78)
	PRANK$_{+F}$	1.62	1.26 (0.78)	5.88	4.57 (0.78)
D500	ORIG	0.46	0.27 (0.58)	1.44	0.76 (0.53)
	MUSCLE	0.44	0.28 (0.63)	1.38	0.84 (0.61)
	MAFFT	0.45	0.30 (0.67)	1.37	0.89 (0.65)
	PRANK$_{+F}$	0.47	0.37 (0.79)	1.41	1.07 (0.76)
D628	ORIG	0.71	1.81 (2.53)	1.66	3.32 (2.00)
	MUSCLE	0.86	2.00 (2.33)	2.03	4.11 (2.02)
	MAFFT	0.36	2.05 (5.64)	0.79	3.27 (4.13)
	PRANK$_{+F}$	0.34	1.61 (4.72)	0.77	2.47 (3.21)
D714	ORIG	0.45	0.36 (0.81)	1.37	0.81 (0.59)
	MUSCLE	0.45	0.48 (1.06)	1.25	0.75 (0.60)
	MAFFT	0.34	0.42 (1.26)	1.22	0.94 (0.77)
	PRANK$_{+F}$	0.42	0.74 (1.79)	1.30	1.31 (1.01)
D855	ORIG	0.59	0.70 (1.19)	0.99	0.87 (0.88)
	MUSCLE	0.63	0.90 (1.43)	1.15	1.20 (1.04)
	MAFFT	0.61	0.51 (0.83)	1.09	0.73 (0.67)
	PRANK$_{+F}$	0.74	1.40 (1.90)	1.33	2.05 (1.54)
D1604	ORIG	0.25	0.90 (3.53)	0.63	0.80 (1.28)
	MUSCLE	0.40	1.03 (2.61)	0.76	0.92 (1.21)
	MAFFT	0.26	0.82 (3.18)	0.65	0.79 (1.21)
	PRANK$_{+F}$	0.34	1.65 (4.80)	0.84	1.39 (1.64)

Table 4. Placement accuracy for the two QS alignment methods on all datasets using QS of length 200 ± 60 bp **with** simulated 454 errors.

```
[MODEL]    GTRexample
    [submodel]  GTR 1.73493250904208
                    0.541209389825999
                    0.316710267113439
                    0.29275303566791
                    0.227817296181108
    [statefreq] 0.180253048026668
                0.254287401144205
                0.234354776256882
                0.331104774572245
    [indelmodel] POW 1.72214718477414 1202
    [indelrate] 0.0002
```

We then used the simulated MSA as input for our evaluation pipline, which means that, we randomly split the MSA into a RA and a set of QS again. Note that, for our evaluation procedure it is necessary to infer a RT for the RA that only contains 50% of the simulated sequences. The RT can be obtained by (i) reconstructing a best-known ML tree on the RA with RAxML or (ii) pruning the QS from the comprehensive true tree on which the data was simulated. We have evaluated both approaches. As in the experiments on real-world data, we sampled short QS of lengths 100 ± 10bp with simulated 454 read errors, and subsequently aligned the short QS against the RA and placed them into the two alternative RTs (best-known ML and pruned true tree). On the simulated MSA using the best-know ML tree, the mean errors (node distances) are 0.12 and 0.24 for PaPaRa and HMMALIGN respectively (0.11 and 0.23 for the pruned RT tree obtained from the true tree). The corresponding normalized edge distances are 0.38% and 0.37% respectively on the best-known ML tree (0.398% and 0.377% on the pruned RT obtained from the true tree). These distances are considerably lower than the distances obtained for the real-world MSAs, despite the fact that it is not entirely clear how to best obtain the RT for the RA. Thus, our experiments on synthetic data indicate that, simultaneuous phylogenetic placement and alignment on real-world data sets constitutes a hard or in other terms sufficiently realistic experimental setting.

5 ONE-SIDED ALIGNMENT ALGORITHM

In an earlier version of PaPaRa, we used a one-sided alignment scheme. In contrast to the current version, this version could not handle insertions in the QS (i.e., delete non-alignable characters from the QS). This version of the algorithm performs well, as long as the QS do not contain insertions with respect to the sequences in the RA. Tables 5 and 6 show the performance of the one-sided

PaPaRa version and HMMALIGN on simulated QS (as described in the main paper) *without* simulated 454 read errors. In this case, the performance differences between the two methods are generally higher (with a maximum difference of a factor of 8.85 for the D1604 data set on the PRANK$_{+F}$ aligned RA) than for the two-sided PaPaRa version. Here the one-sided PaPaRa version can take advantage of the rather unrealistic assumption that, the QS do not contain insertions, which we make for our simulated QS *without* simulated 454 sequencing errors. On the QS *with* simulated errors it performs worse than HMMALIGN (results not shown). The one-sided version has lower runtimes than the two-sided version (faster by approximately a factor of 3) because of the lower algorithmic complexity in the alignment core function and an additional early-stopping rule that can be used to stop the calculation of the dynamic programming matrix early.

		ND		EDN %	
D	MA	PaPaRa	HMM	PaPaRa	HMM
D150	ORIG	0.17	0.60 (3.65)	0.48	0.98 (2.03)
	MUSCLE	0.16	0.49 (3.01)	0.52	0.60 (1.14)
	MAFFT	0.15	0.56 (3.63)	0.54	0.78 (1.44)
	PRANK$_{+F}$	0.23	1.03 (4.40)	0.69	1.42 (2.05)
D218	ORIG	1.03	1.07 (1.04)	3.50	3.61 (1.03)
	MUSCLE	1.08	0.93 (0.87)	4.47	3.96 (0.89)
	MAFFT	1.05	1.17 (1.12)	3.62	4.11 (1.13)
	PRANK$_{+F}$	0.94	1.25 (1.33)	3.43	4.49 (1.31)
D500	ORIG	0.09	0.18 (1.99)	0.25	0.49 (1.93)
	MUSCLE	0.08	0.19 (2.30)	0.24	0.57 (2.40)
	MAFFT	0.08	0.21 (2.70)	0.26	0.63 (2.45)
	PRANK$_{+F}$	0.17	0.31 (1.84)	0.44	0.84 (1.90)
D628	ORIG	0.62	1.66 (2.67)	1.59	3.00 (1.88)
	MUSCLE	0.79	1.91 (2.42)	1.87	3.88 (2.08)
	MAFFT	0.31	1.98 (6.45)	0.75	3.13 (4.15)
	PRANK$_{+F}$	0.34	1.50 (4.45)	0.73	2.33 (3.22)
D714	ORIG	0.19	0.30 (1.56)	0.48	0.65 (1.35)
	MUSCLE	0.21	0.42 (2.03)	0.45	0.64 (1.42)
	MAFFT	0.14	0.36 (2.55)	0.42	0.78 (1.88)
	PRANK$_{+F}$	0.20	0.70 (3.50)	0.50	1.20 (2.39)
D855	ORIG	0.24	0.55 (2.31)	0.39	0.72 (1.83)
	MUSCLE	0.33	0.74 (2.27)	0.54	0.98 (1.82)
	MAFFT	0.26	0.41 (1.57)	0.43	0.58 (1.36)
	PRANK$_{+F}$	0.37	1.25 (3.39)	0.62	1.85 (2.98)
D1604	ORIG	0.09	0.80 (8.74)	0.20	0.68 (3.40)
	MUSCLE	0.25	0.95 (3.82)	0.37	0.81 (2.21)
	MAFFT	0.10	0.72 (7.14)	0.21	0.69 (3.30)
	PRANK$_{+F}$	0.18	1.55 (8.85)	0.32	1.30 (4.07)

Table 5. Placement accuracy for the one-sided version of PaPaRa on all data-sets using QS of length 200 ± 60 bp **without** simulated 454 errors.

REFERENCES

Balzer, S., Malde, K., Lanzn, A., Sharma, A., and Jonassen, I. (2010). Characteristics of 454 pyrosequencing dataenabling realistic simulation with flowsim. *Bioinformatics*, **26**(18), i420–i425.

Cartwright, R. (2005). DNA Assembly With Gaps (Dawg): Simulating Sequence Evolution. *Bioinformatics*, **21**(Suppl. 3), iii31–iii38.

		ND		EDN %	
D	MA	PaPaRa	HMM	PaPaRa	HMM
D150	ORIG	0.14	0.17 (1.23)	0.38	0.57 (1.52)
	MUSCLE	0.12	0.11 (0.92)	0.52	0.22 (0.41)
	MAFFT	0.14	0.19 (1.39)	0.52	0.38 (0.73)
	PRANK$_{+F}$	0.19	0.37 (1.88)	0.58	0.56 (0.96)
D218	ORIG	0.65	0.55 (0.84)	2.34	1.83 (0.78)
	MUSCLE	0.77	0.43 (0.57)	3.48	1.94 (0.56)
	MAFFT	0.69	0.67 (0.96)	2.61	2.62 (1.01)
	PRANK$_{+F}$	0.55	0.77 (1.41)	2.08	2.89 (1.39)
D500	ORIG	0.03	0.03 (1.17)	0.13	0.14 (1.06)
	MUSCLE	0.03	0.03 (1.34)	0.11	0.13 (1.20)
	MAFFT	0.03	0.04 (1.41)	0.16	0.19 (1.17)
	PRANK$_{+F}$	0.04	0.05 (1.28)	0.17	0.20 (1.20)
D628	ORIG	0.56	1.49 (2.68)	1.08	2.30 (2.14)
	MUSCLE	0.81	1.17 (1.44)	1.65	2.02 (1.23)
	MAFFT	0.18	1.04 (5.76)	0.24	1.51 (6.31)
	PRANK$_{+F}$	0.21	0.98 (4.73)	0.39	1.04 (2.67)
D714	ORIG	0.12	0.16 (1.32)	0.32	0.33 (1.02)
	MUSCLE	0.15	0.17 (1.12)	0.30	0.27 (0.91)
	MAFFT	0.09	0.13 (1.50)	0.24	0.27 (1.09)
	PRANK$_{+F}$	0.14	0.29 (2.03)	0.28	0.47 (1.70)
D855	ORIG	0.11	0.17 (1.61)	0.19	0.22 (1.17)
	MUSCLE	0.16	0.24 (1.53)	0.30	0.29 (0.99)
	MAFFT	0.13	0.18 (1.35)	0.24	0.28 (1.17)
	PRANK$_{+F}$	0.16	0.38 (2.38)	0.30	0.44 (1.49)
D1604	ORIG	0.08	0.35 (4.50)	0.14	0.29 (2.07)
	MUSCLE	0.26	0.54 (2.06)	0.30	0.30 (0.99)
	MAFFT	0.07	0.36 (5.33)	0.14	0.27 (1.94)
	PRANK$_{+F}$	0.11	0.73 (6.40)	0.21	0.53 (2.49)

Table 6. Placement accuracy for the one-sided (OS) version of PaPaRa on all data-sets using QS of length 500 ± 60 bp **without** simulated 454 errors.

Fletcher, W. and Yang, Z. (2009). INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. and Evol.*, **26**(8), 1879 – 1888.

Nethercote, N. and Seward, J. (2007). Valgrind: A Framework for Heavyweight Dynamic Binary Instrumentation. In *Proc. of PLDI 2007*.

Stamatakis, A., Komornik, Z., and Berger, S. A. (2010). Evolutionary placement of short sequence reads on multi-core architectures. In *Proc. of AICCSA-10*. accepted for publication.