Supplemental Table S1 *k*-mer analysis of PacBio reads.

PacBio dataset	Genome size (Mb)	Coverage*	Reference *k*-mer list†	Total *k*-mers				Distinct *k*-mers			
				N	Valid	Not valid	% valid ‡	N	Valid	Not valid	% valid
E. coli	4.64	94	genome	435753159	66902393	368850766	15.35	292687635	4513248	288174387	1.54
E. coli	4.64	94	Illumina	435753159	67632607	368120552	15.52	292687635	4545559	288142076	1.55
S. cerevisae	12.1	116	Illumina	1399080968	236336855	1162744113	16.89	656283404	11312654	644970750	1.72
C. elegans	103	79	genome	8106551865	3027061318	5079490547	37.34	1453998486	73682446	1380316040	5.07
C. elegans	103	79	Illumina	8106551865	3111708811	4994843054	38.39	1453998486	79509922	1374488564	5.47
Arabidopsis	135	137	Illumina	18488719952	6596898991	11891820961	35.68	1838858368	108591962	1730266406	5.91
Drosophila	180	87	Illumina	15718720226	4796884965	10921835261	30.52	1930342727	107187632	1823155095	5.55
Random sequence	100	100	synthetic genome	11091044996	1421548773	9669496223	12.82	2123282004	97708253	2025573751	4.60

* Coverage estimated with jellyfish (Marçais and Kingsford 2011) as total *k*-mers / genome size

† Source of the valid *k*-mer list. "genome" means all *k*-mers from the finished genome is used; "Illumina", *k*-mers from Illumina reads filtered from low-frequency *k*-mers; "synthetic genome", all *k*-mers from the random sequence used to generate the synthetic PacBio reads.

‡ Repeat-rich genomes have a higher proportion of correct *k*-mers possibly because errors in repetitive sequences have a rather high chance of generating a valid *k*-mer that occurs in a variant copy of the repeat (located elsewhere in the genome).