Real-time resolution of short-read assembly graph using ONT long reads

Son Hoang Nguyen[1*], Minh Duc Cao[1], Lachlan Coin[1,2,3,4*]

1 Institute for Molecular Bioscience, the University of Queensland, St Lucia, Brisbane, QLD 4072 Australia
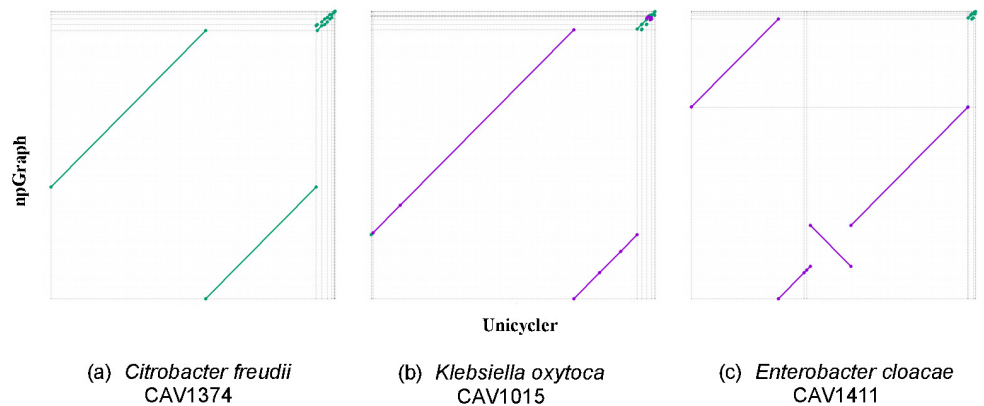**2** Department of Microbiology and Immunology, The University of Melbourne, Parkville, 3010, Australia
3 Department of Clinical Pathology, The University of Melbourne, Parkville, 3010, Australia
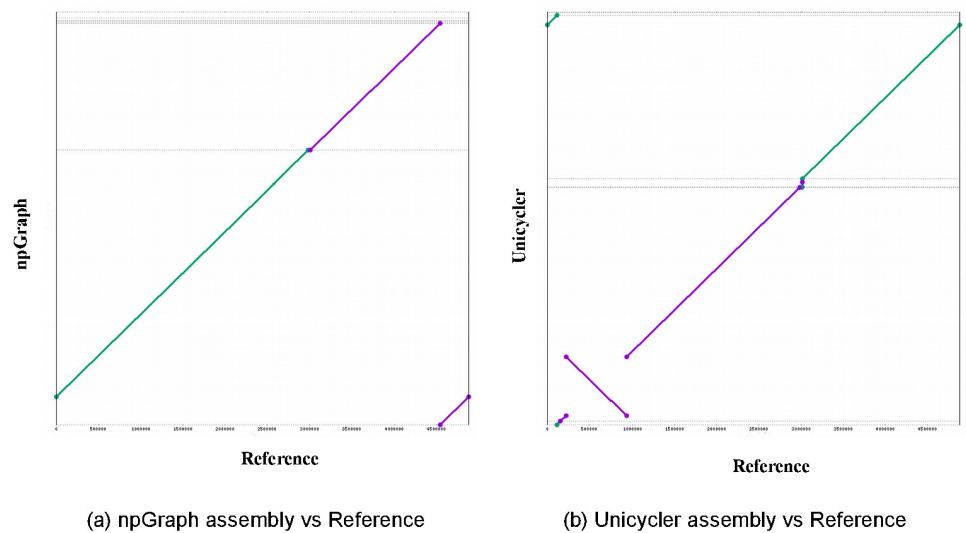**4** Department of Infectious Disease, Imperial College London, London, W2 1NY, UK

*s.hoangnguyen@imb.uq.edu.au,l.coin@imb.uq.edu.au

Supporting information

(a) *Citrobacter freudii* CAV1374

(b) *Klebsiella oxytoca* CAV1015

(c) *Enterobacter cloacae* CAV1411

S1 Fig. Dotplot generated by MUMmer for assembly results of `Unicycler` versus `npGraph`. Structural agreements between two methods were found in (a) *C.freundii* and (b) *K.oxytoca* assembly contigs. On the other hand, for (c) *E.cloacae* sample, there was a disagreement detected between 2 largest contigs given by two assembly algorithms.

(a) npGraph assembly vs Reference

(b) Unicycler assembly vs Reference

S2 Fig. Alignments of an *Enterobacter cloacae* reference genome to assembly sequences generated by (a) `npGraph` and (b) `Unicycler`. The former suggests a structural variant, the latter is virtually an 1-to-1 mapping.

S1 Table Benchmarking different methods using `LRScaf`, `npScarf`, `npGraph`, `hybridSPAdes` and `Unicycler` hybrid assembler with the synthetic data set.

Method	Assembly size (bp)	#Contigs	N50 (bp)	Mis-assemblies	Mismatch (per 100Kbp)	Indel (per 100Kbp)
random sequences no repeats						
LRScaf	-	-	-	-	-	-
npScarf	4110000	3	4000000	0	0.00	0.00
npScarf_wag	4109516	3	4000000	0	0.00	0.00
npGraph-bwa	4110000	3	4000000	0	0.00	0.00
npGraph-mm2	4110000	3	4000000	0	0.00	0.00
hybridSPAdes	4110231	3	4000077	0	0.00	0.07
Unicycler	4110000	3	4000000	0	0.00	0.00
random sequences some repeats						
LRScaf	4113104	6	4002298	0	0.00	0.00
npScarf	4110940	3	4002715	0	0.00	0.66
npScarf_wag	4437094	7	2795129	0	0.00	1.07
npGraph-bwa	4110000	3	4000000	0	0.00	0.00
npGraph-mm2	4110000	3	4000000	0	0.00	0.00
hybridSPAdes	4107566	3	3999364	0	0.02	0.02
Unicycler	4110000	3	4000000	0	0.00	0.00
random sequences many repeats						
LRScaf	4119020	12	4002117	0	0.03	0.79
npScarf	4316934	8	3963485	24	11.55	34.54
npScarf_wag	5215965	16	1515563	37	0.32	7.22
npGraph-bwa	4110000	3	4000000	0	0.32	0.15
npGraph-mm2	4110000	3	4000000	0	0.32	0.15
hybridSPAdes	4108190	3	3999621	0	0.68	0.15
Unicycler	4110000	3	4000000	0	0.32	0.15
Acinetobacter A1						
LRScaf	3896799	4	3886381	0	0.47	0.36
npScarf	3912299	3	3870269	4	4.74	13.02
npScarf_wag	3945166	3	3906368	1	7.00	14.02
npGraph-bwa	3918374	2	3909643	0	16.34	0.61
npGraph-mm2	3885898	2	3877167	1	18.05	1.03
hybridSPAdes	3929948	53	3353679	0	35.48	3.22
Unicycler	3917745	2	3909014	0	2.50	0.13
Acinetobacter AB30						
LRScaf	4367657	27	1035967	13	30.11	1.42
npScarf	4512464	7	4304628	35	57.95	72.87
npScarf_wag	5315235	13	1267710	136	73.55	8.15
npGraph-bwa	4335342	2	4148952	1	16.93	1.45
npGraph-mm2	4335790	1	4335790	0	6.97	0.25
hybridSPAdes	4337369	3	2701005	0	12.80	1.39
Unicycler	4333041	1	4333041	1	6.42	0.53

Table 1 – *Continued from previous page*

Method	Assembly size (bp)	#Contigs	N50 (bp)	Mis-assemblies	Mismatch (per 100Kbp)	Indel (per 100Kbp)
E. coli K12 MG1655						
LRScaf	4732001	16	692794	7	3.42	0.58
npScarf	4649902	2	4641702	4	14.94	34.35
npScarf_wag	4687952	3	4641732	0	6.55	1.96
npGraph-bwa	4641743	1	4641743	0	4.50	0.43
npGraph-mm2	4641820	1	4641820	0	3.88	0.26
hybridSPAdes	4644555	1	4641036	0	0.62	0.09
Unicycler	4641650	1	4641650	0	3.43	0.26
E. coli O25b H4-ST131						
LRScaf	5410854	21	910949	9	13.70	0.87
npScarf	5245913	3	5095571	7	7.05	18.81
npScarf_wag	5292700	3	3469617	9	9.03	1.55
npGraph-bwa	5237821	7	4049493	1	3.38	0.31
npGraph-mm2	5249799	3	5110117	0	2.40	0.15
hybridSPAdes	5252762	8	4258948	2	5.43	0.57
Unicycler	5249442	3	5109760	0	4.02	0.27
Klebsiella 30660 NJST258 1						
LRScaf	5512518	20	777919	26	3.52	0.38
npScarf	5559772	7	5259053	4	17.18	13.48
npScarf_wag	5613780	7	5268535	6	1.59	1.92
npGraph-bwa	5534843	8	5263229	0	3.15	0.76
npGraph-mm2	5534878	8	5263264	0	2.75	0.74
hybridSPAdes	5545668	8	5545668	2	4.95	0.09
Unicycler	5537860	9	5263196	0	1.34	0.51
Klebsiella MGH 78578						
LRScaf	5823340	38	559414	7	5.92	1.27
npScarf	5729304	5	5316429	12	14.90	20.27
npScarf_wag	5754443	5	3026286	16	8.17	2.56
npGraph-bwa	5695801	7	5311745	1	12.65	1.25
npGraph-mm2	5696302	6	5315267	0	6.06	0.44
hybridSPAdes	5706470	11	5315273	1	3.82	0.67
Unicycler	5694231	14	5315096	0	5.38	0.21
Klebsiella NTUH-K2044						
LRScaf	5470282	8	2951488	9	3.04	2.62
npScarf	5471696	2	5249198	6	4.82	8.55
npScarf_wag	5530559	3	5249369	2	2.25	1.35
npGraph-bwa	5472629	2	5248476	0	2.52	0.31
npGraph-mm2	5472655	2	5248503	0	1.21	0.26
hybridSPAdes	5473572	2	5248894	0	0.44	0.15
Unicycler	5472697	2	5248545	0	2.41	0.35
Mycobacterium tuberculosis H37Rv						
LRScaf	4260869	12	3215170	8	6.69	2.27
npScarf	4498245	4	4402238	8	5.15	2.68
npScarf_wag	4506056	4	4410942	3	6.81	2.59
npGraph-bwa	4411406	1	4411406	0	1.88	0.43
npGraph-mm2	4411532	1	4411532	0	0.68	0.00
hybridSPAdes	4413942	1	4410519	0	0.75	0.11
Unicycler	4411538	1	4411538	0	2.22	0.34

Table 1 – *Continued from previous page*

Method	Assembly size (bp)	#Contigs	N50 (bp)	Mis-assemblies	Mismatch (per 100Kbp)	Indel (per 100Kbp)
Saccharomyces cerevisiae S288c						
LRScaf	12486884	214	254587	27	16.45	1.97
npScarf	11986800	24	796769	51	62.12	21.46
npScarf_wag	12003203	21	917017	21	69.14	5.47
npGraph-bwa	11921736	40	913090	3	38.04	1.94
npGraph-mm2	11920984	38	913198	2	20.66	0.95
hybridSPAdes	12027533	45	770543	5	32.58	1.94
Unicycler	11847655	72	909114	0	21.81	1.04
Shigella dysenteriae Sd197						
LRScaf	4678609	175	376897	26	3.91	2.35
npScarf	4586075	173	36560	55	120.14	111.59
npScarf_wag	5462918	92	98791	105	147.48	79.28
npGraph-bwa	4564058	6	4369264	5	80.64	11.16
npGraph-mm2	4558920	6	4364264	7	75.51	10.98
hybridSPAdes	4519131	23	821249	96	9.57	1.42
Unicycler	4560901	3	4369231	0	11.88	1.05
Shigella sonnei 53G						
LRScaf	5266645	128	429353	27	9.72	1.19
npScarf	6441461	20	1953896	82	164.02	219.52
npScarf_wag	-	-	-	-	-	-
npGraph-bwa	5211544	4	4988532	0	14.53	0.31
npGraph-mm2	5211527	4	4988519	0	8.56	0.17
hybridSPAdes	5223875	8	2195455	2	41.92	0.06
Unicycler	5220517	5	4988548	0	7.39	0.52
Streptococcus suis BM407						
LRScaf	2231488	10	683404	6	11.96	1.37
npScarf	2183951	3	2146594	0	21.51	9.17
npScarf_wag	2289880	3	1493189	1	3.17	1.96
npGraph-bwa	2154623	6	2131479	1	5.25	0.28
npGraph-mm2	2149876	6	2146774	0	1.44	0.28
hybridSPAdes	2172703	2	2146237	0	6.82	0.09
Unicycler	2170829	2	2146250	0	2.67	0.32