Supplementary material for the manuscript:

In-silico read normalization using set multi-cover optimization

Dilip A Durai and Marcel H Schulz

1 Parameters used for the evaluation

Two well established normalization methods- Diginorm and Trinity's in-silico normalization (TIS) were used for comparing and evaluating the performance of ORNA. Given a coverage threshold c, Diginorm streams through the read dataset. For each read, it calculates the median abundance of k-mers present in the read. The abundance information is obtained from the previously accepted reads. Once the median abundance goes beyond c, the read is rejected.

Trinity's in-silico normalization, on the other hand, pre-calculates the k-mer abundance in each read. It then iterates over the dataset. For a desired threshold c, TIS calculates the median abundance of the k-mers in the read. If the median coverage is less than the desired coverage, the read is always kept. If the median coverage is greater than the desired coverage, a random number is generated between 0 and 1. If this number is less than the ratio of the desired coverage and the median coverage, the read is kept otherwise it is removed.

Table S1: Parameters used for different normalization algorithms. ORNA requires base b of the logarithm function. Diginorm and TIS requires a coverage cut-off value c. The first column depicts the k-mer size used for the normalization algorithms. The second column contains the datasets used for the algorithms. A star next to the value indicates the default coverage/base of the algorithm.

k-mer size	Dataset	ORNA(b)	Diginorm(c)	TIS(c)
22	SRR332171	(1.3,1.5,1.7*,3,5,7,10)	(5,10*,15,20,25,30)	(5,10,15,20,25,30)
	SRR1020625	(1.3,1.5,1.7*,3,5,7,10)	(15,20,25,30,35,45,50,55,60,65,70)	(30,35,45,50*,55,60,65,70)
	combined	5	10	10
26	SRR332171	(1.3,1.5,1.7*,3,5,7,9)	(5,10*,15,20,25,30)	(5,10,15,20,25,30)
	SRR1020625	(1.7*,3,5,7,9,15,25 ,35,100,200,300)	(1,5,10*,20,30,40,50,60,70,80)	(5,10,20,30,40,50*, 60,70,80,90,100,110)

Table S2: Parameters used for different normalization algorithms in paired-end mode. ORNA requires base b of the logarithm function. Diginorm and TIS requires a coverage cut-off value c. The first column depicts the k-mer size used for the normalization algorithms. The second column contains the dataset used for the algorithms.

k-mer size	Dataset	ORNA(b)	Diginorm(c)	TIS(c)
22	SRR332171	(1.3,1.5,1.7*,3,5,7)	(3,5,10*,15,20,25)	(10,15,20,25,30,35,40)

Table S3: Other non-default parameters used for the Diginorm and TIS

Diginorm	TIS
–hashsize 32e+9	–JM 10G
–n_hashes 4	–max_pct_stdev 10000

2 Assembly performance of normalization algorithms

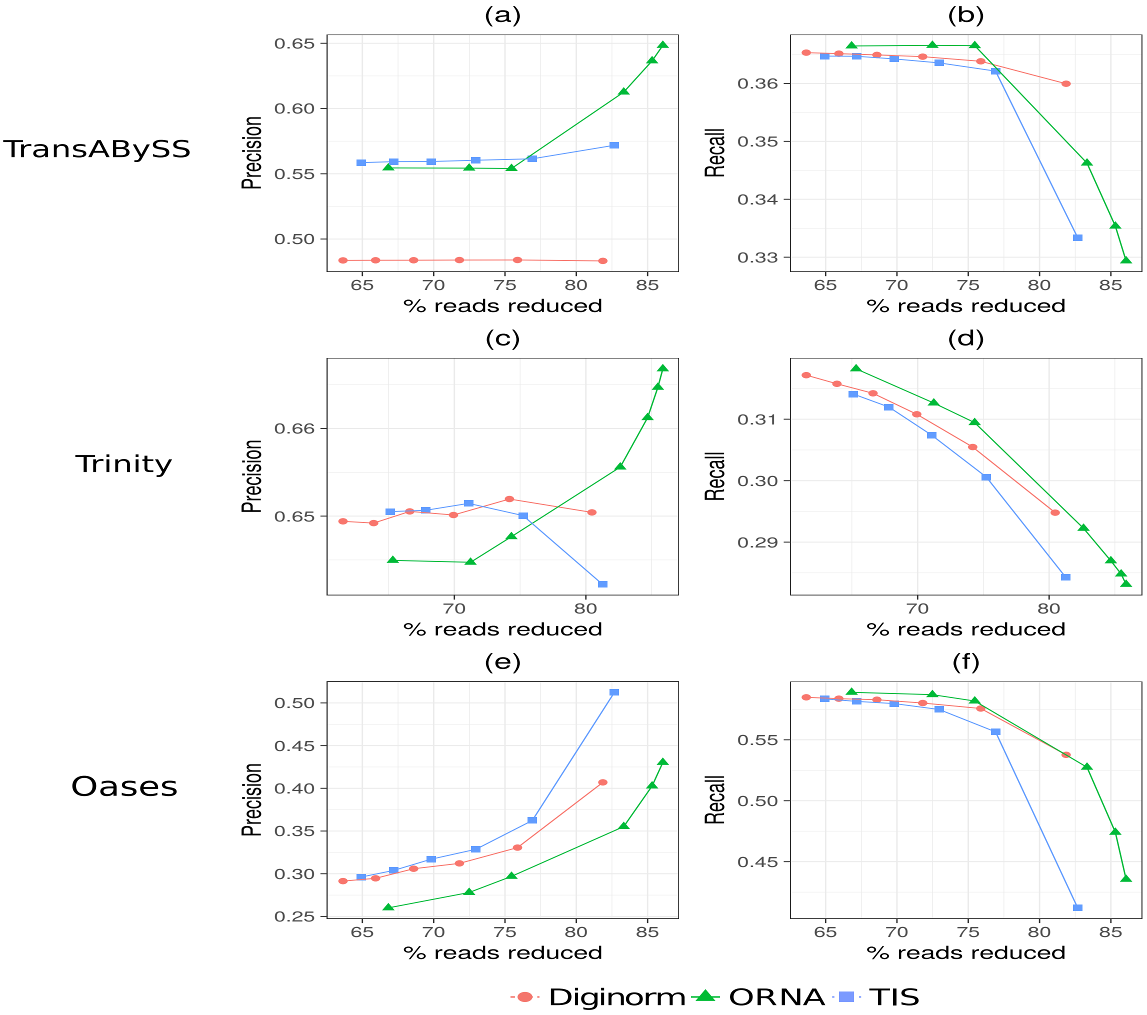

Figure S1: Comparison of precision (a,c,e) and recall (b,d,f) scores obtained from assemblies of ORNA, Diginorm and TIS reduced brain datasets. Each point on a line corresponds to a different parametrization of the algorithms. The amount of data reduction (x-axis) is compared against the precision/recall measured from REF-EVAL (y-axis, see main text). (a) and (b) represent TransABySS assemblies (k=21). (c) and (d) represent Trinity assemblies (k=25) and (e) and (f) represent Oases multi-$k$mer assemblies applied on normalized brain.

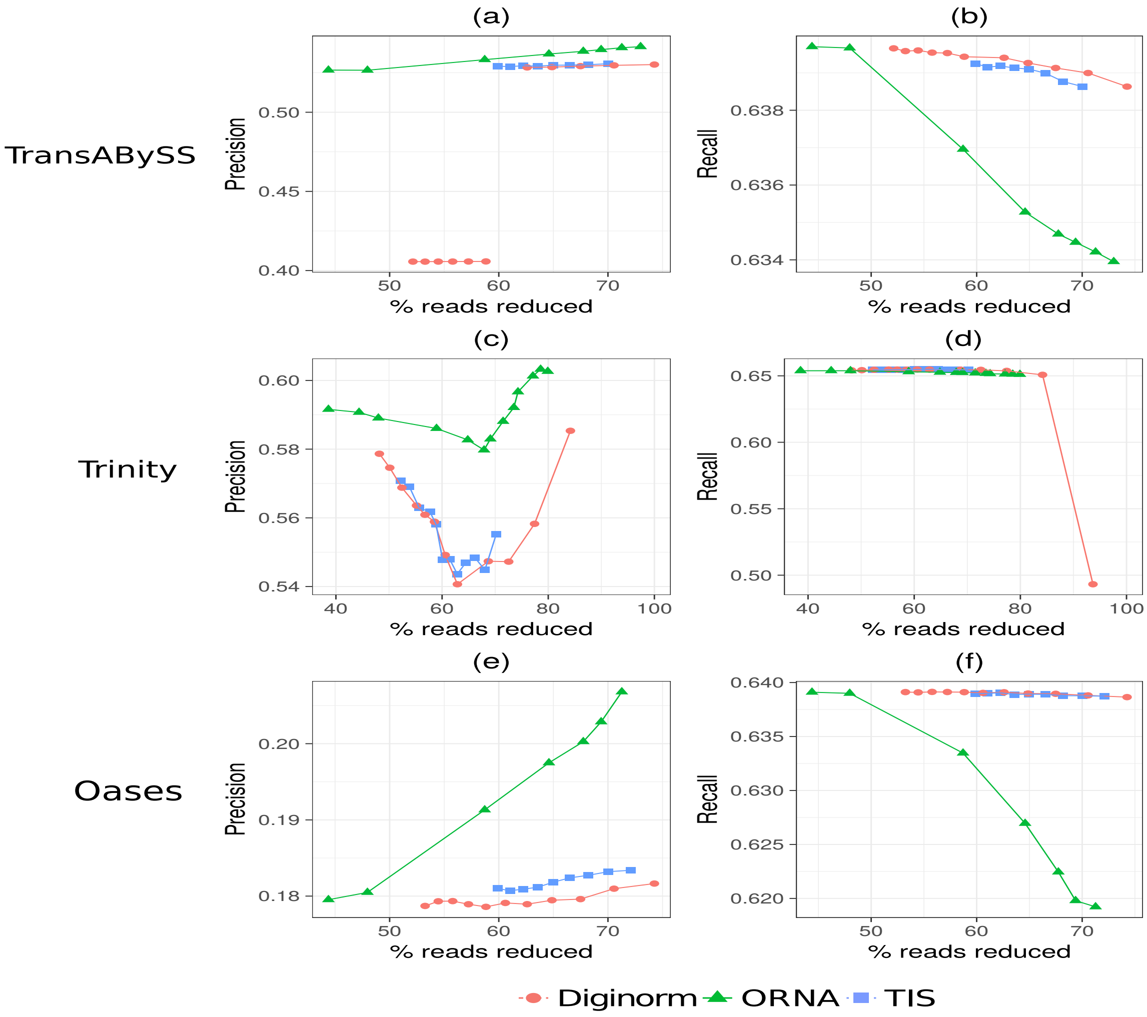

Figure S2: Comparison of precision (a,c,e) and recall (b,d,f) scores obtained from assemblies of ORNA, Diginorm and TIS reduced hESC datasets. Each point on a line corresponds to a different parametrization of the algorithms. The amount of data reduction (x-axis) is compared against the precision/recall measured from REF-EVAL (y-axis, see main text). (a) and (b) represent TransABySS assemblies (k=21). (c) and (d) represent Trinity assemblies (k=25) and (e) and (f) represent Oases multi-$k$mer assemblies applied on normalized brain.

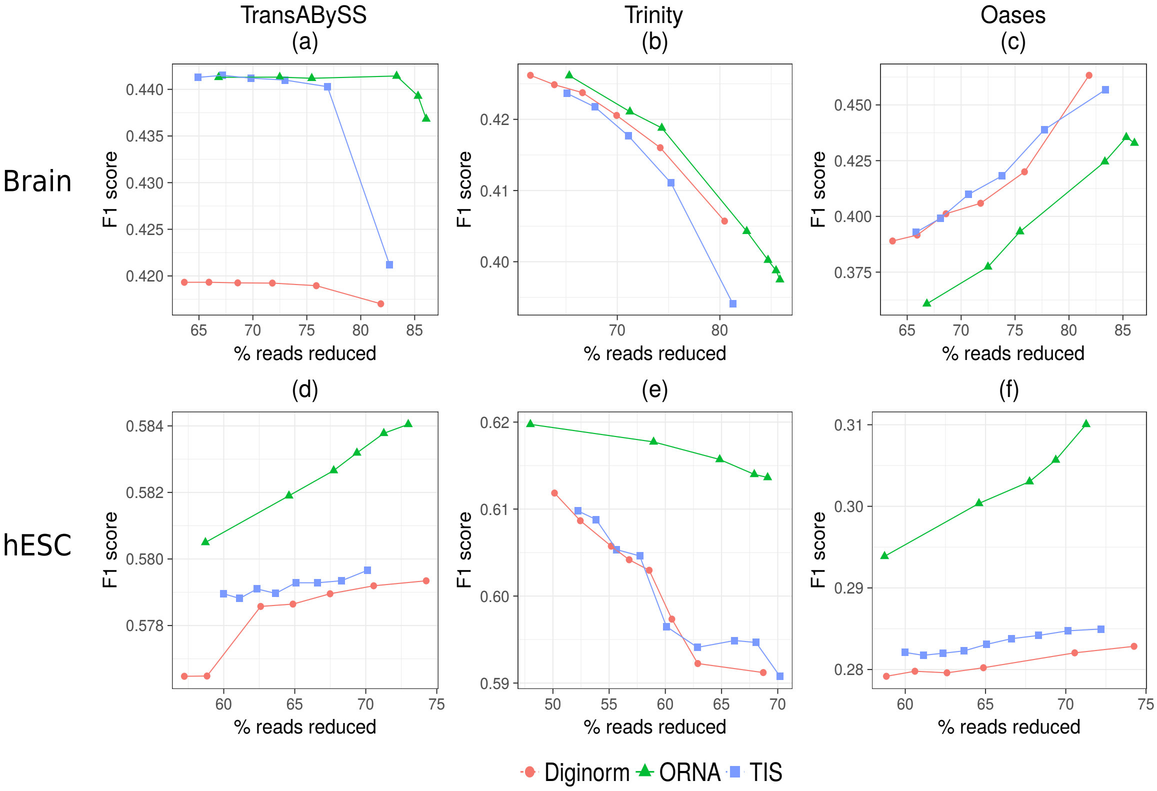

Figure S3: Comparison of F1 scores obtained from assemblies of ORNA, Diginorm and TIS reduced datasets. Each point on a line corresponds to a different parametrization of the algorithms. The amount of data reduction (x-axis) is compared against the F1 score measured from DETONATE (y-axis, see text). (a) and (d) represent TransABySS assemblies (k=21) applied on normalized brain and hESC data, respectively. (b) and (e) represent Trinity assemblies (k=25) and (c) and (f) represent Oases multi-$k$mer assemblies applied on normalized brain and hESC data, respectively.

3 Assembly performance of normalization algorithms in paired-end mode

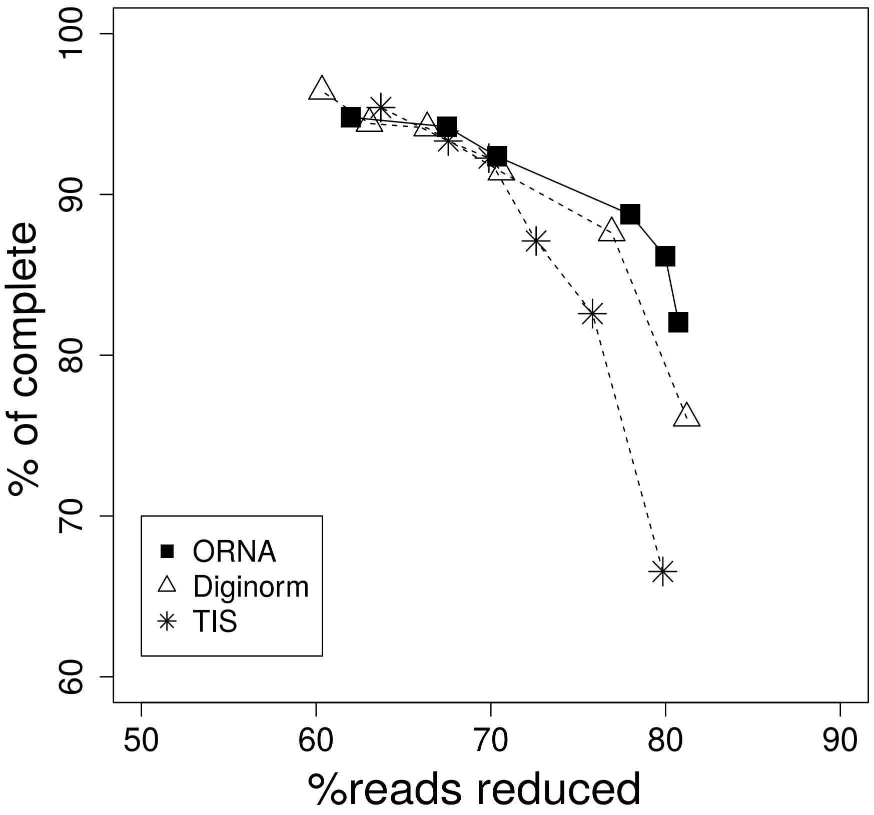

Figure S4: Comparison of TransABySS assemblies generated from ORNA, Diginorm and TIS reduced brain datasets. The normalization algorithms were run in paired-end mode. Each point on a line corresponds to a different parametrization of the algorithms. The amount of data reduction (x-axis) is compared against the assembly performance measured as *% of complete* (y-axis).

4 Memory and runtime requirements for normalization algorithms in paired-end mode

Table S4: Runtime (in minutes) and memory (in GB) required by different normalization algorithms in paired-end mode. Note that the memory of Diginorm can be set by the user. For comparison it is set such that it uses less or similar memory than ORNA denoted as Diginorm ^{a} and ^{b}, respectively. The percent of reads reduced by each method (% reduced) is shown in the first column for each dataset. The total number of reads (in millions) and the file size (in GB) of the original dataset is shown in brackets next to the dataset.

method	brain (147M - 20.1GB)		
	%reduced	time [min]	mem [GB]
ORNA	67.47	135	7.13
Diginorm^a	66.86	195	4.1
Diginorm^b	66.36	213	6.26
TIS	67.57	159	21.79