

Supplemental Table S2. Illumina *k*-mer frequency cut-offs used to produce the valid *k*-mer lists.

Organism	Strain	Genome size (Mbp)	Quality cut-off	Illumina raw coverage *	Illumina net coverage†	Low freq. cut-off ‡	High freq. cut-off §
<i>E. coli</i>	K-12 MG1655	4.64	Q0	77	70	10	105
<i>S. cerevisiae</i>	W303	12.1	Q20	232	189	27	280
<i>C. elegans</i>	Bristol N2	103	Q20	800	586	100	863
<i>Drosophila</i>	ISO1	180	Q10	77 ¶	46 (XY) 93 (autosomes)	13	150
<i>Arabidopsis</i>	Ler-0 (PacBio) Ler-1, Ler-0 (Illumina)	135	Q0	106	40	6	60
<i>H. sapiens</i>	GIAB HG002	3,000	Q0	278	144 (XY) 262 (autosomes)	76	393

* Calculated as Total *k*-mers / genome size. This value is larger than the net coverage because it includes the error *k*-mers. The accession numbers of the reads are in Supplemental Table S12.

† Calculated as peak of single-copy *k*-mer freq. *Drosophila* and human data came from the heterogametic sex, and in these cases the sex chromosomes form a separate peak, with half of the autosomal coverage.

‡ Intended to remove error *k*-mers. Set to approximately one seventh of the net coverage (single-copy peak)

§ Intended to remove *k*-mers with more than one copy in the genome. Set to approximately 1.5 fold the net coverage (single-copy peak).

¶ This value is the composite of two distributions: autosomes and sex-chromosomes (see Supplemental Fig. S5).