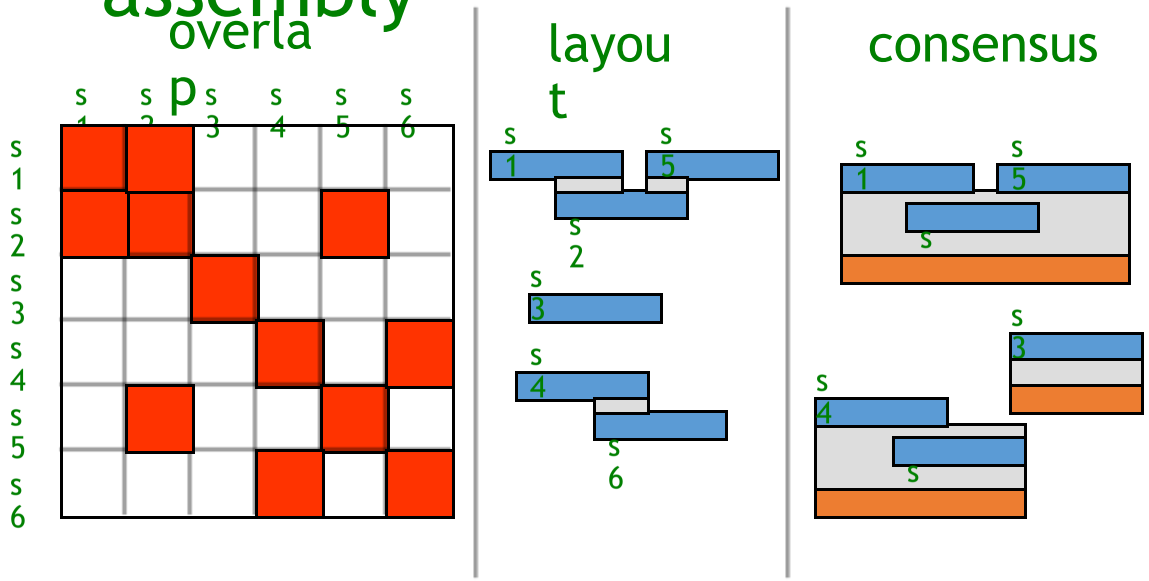
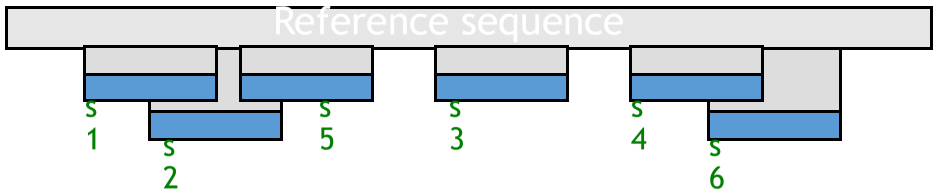
Sequence assembly

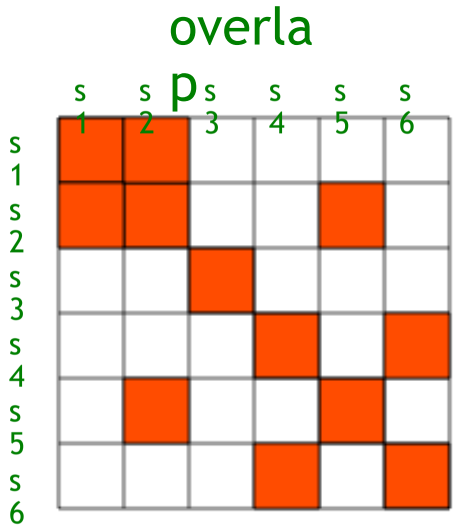
de novo



reference-guided



de novo sequence assembly



Most CPU and memory demanding stage

Phrap: “banded” alignment of reads around k-mer matches; tolerate alignment mismatches of low-quality bases

Phusion: group reads sharing >= 11 k-mers of 17 bases

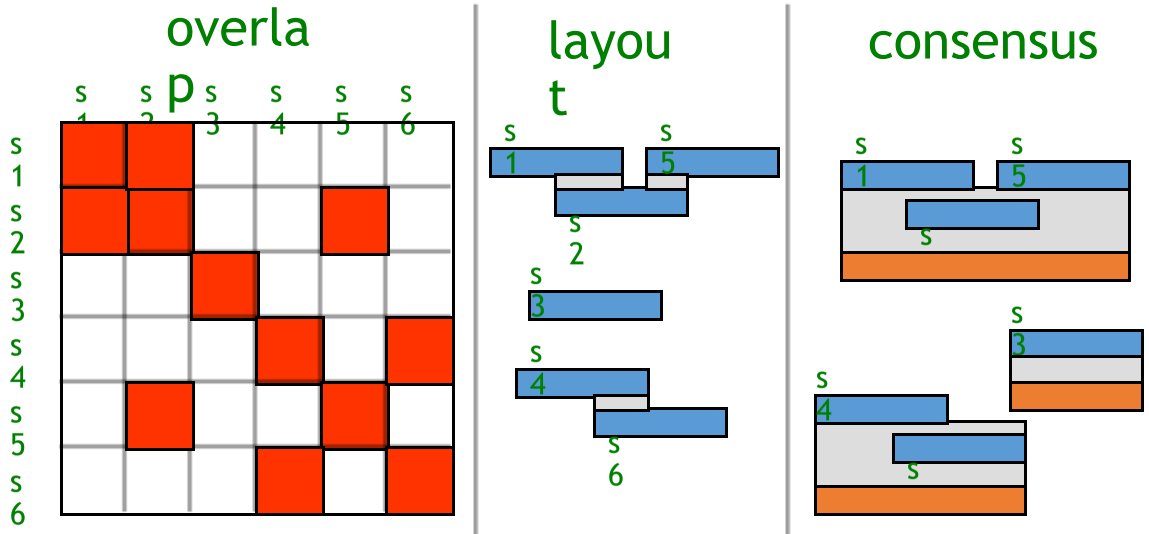
Celera: k-mer seed and extend alignment of reads

Arachne: 24-mer seed and extend alignment of reads

newbler: flowgram similarities (?)

Read 1:AACGTAGCTAC
Read 2:AGCTACGG
Read 3:CGTAGCT

O-L-C



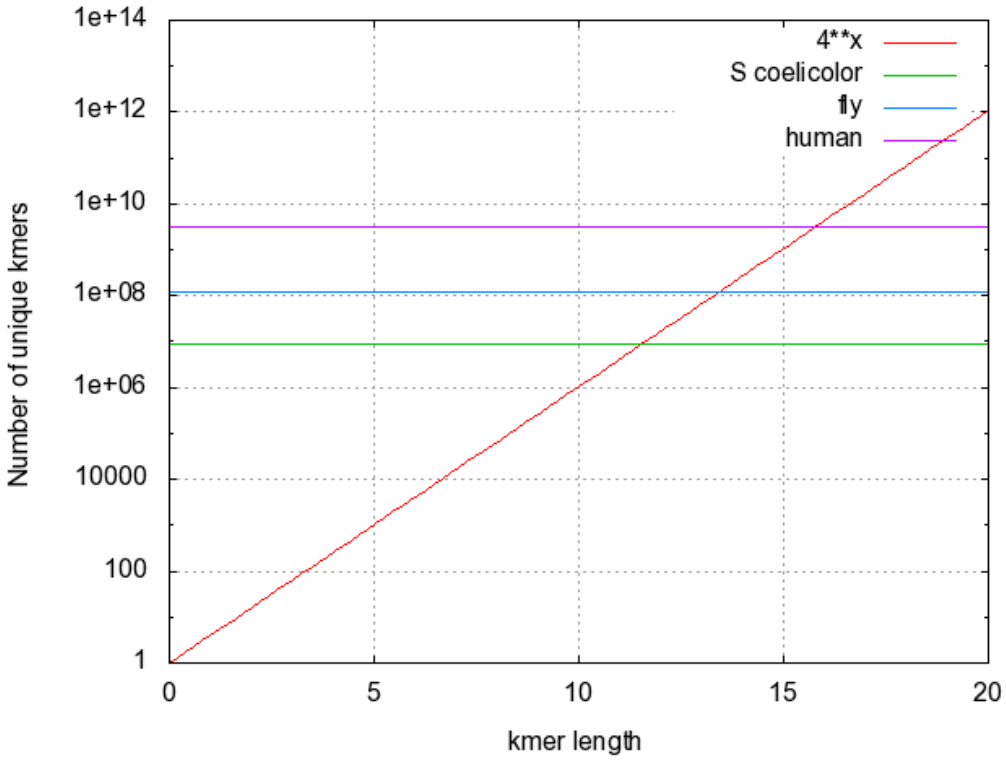
$O(N$

De Bruijn graph

Theoretical genome uniqueness

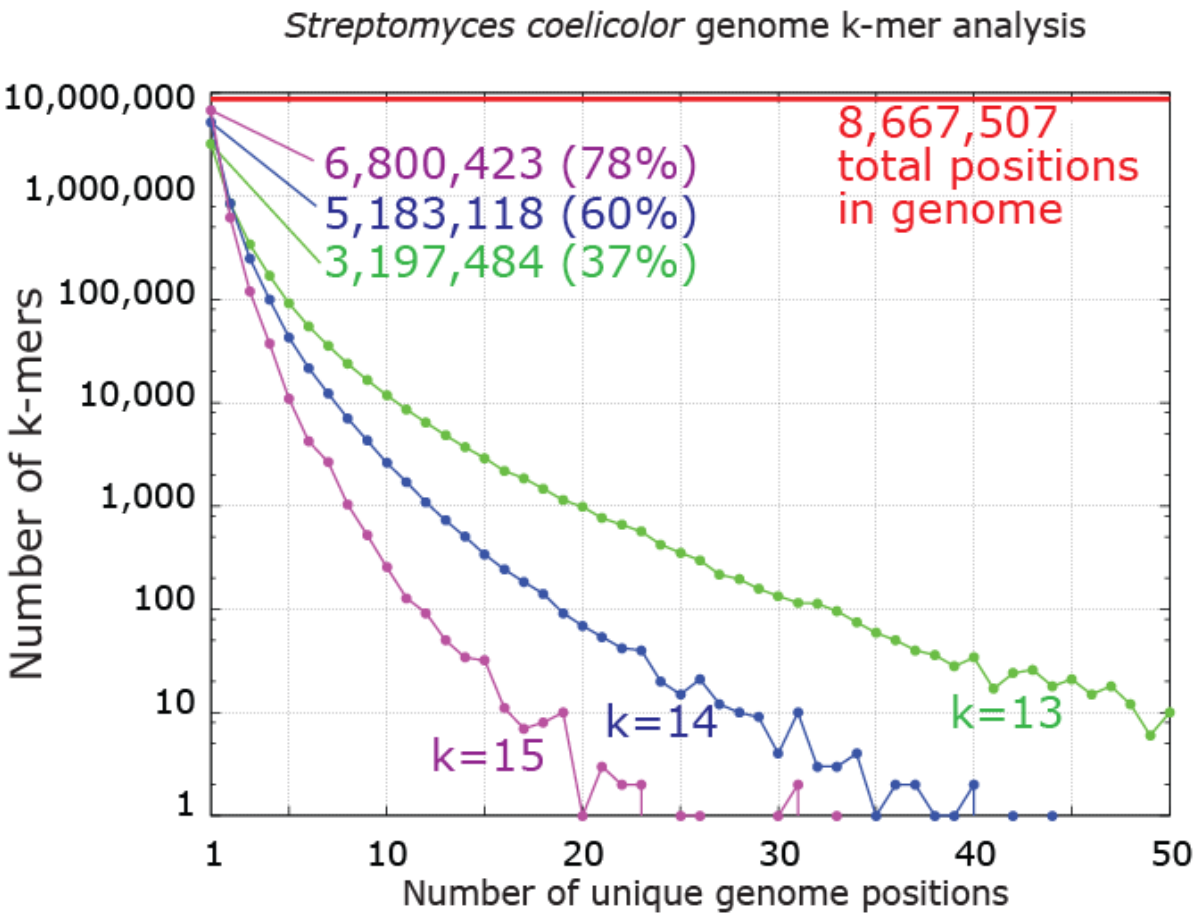
K-mers:
words of length *k*

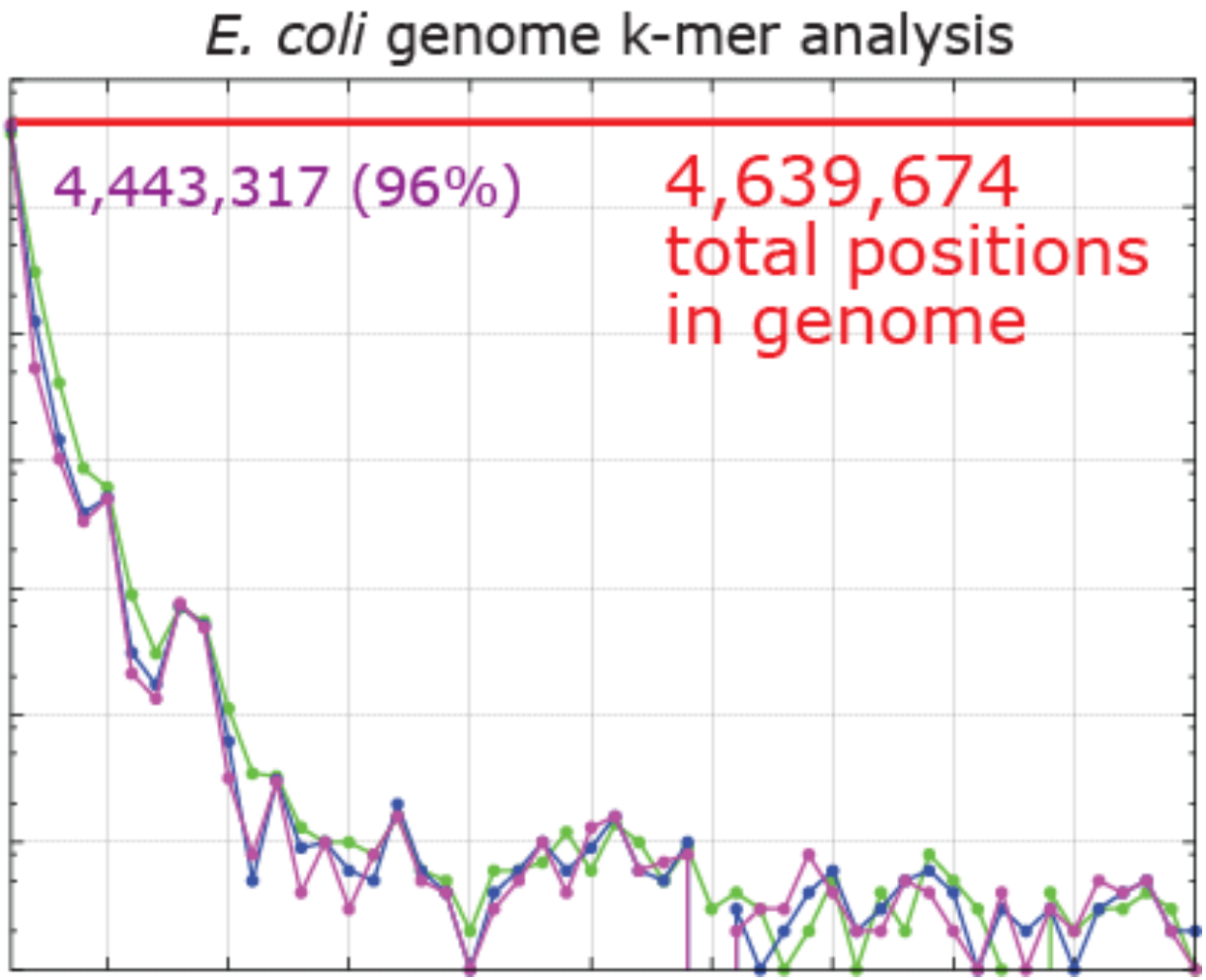
How many *k-mers* of length *n?*
4n



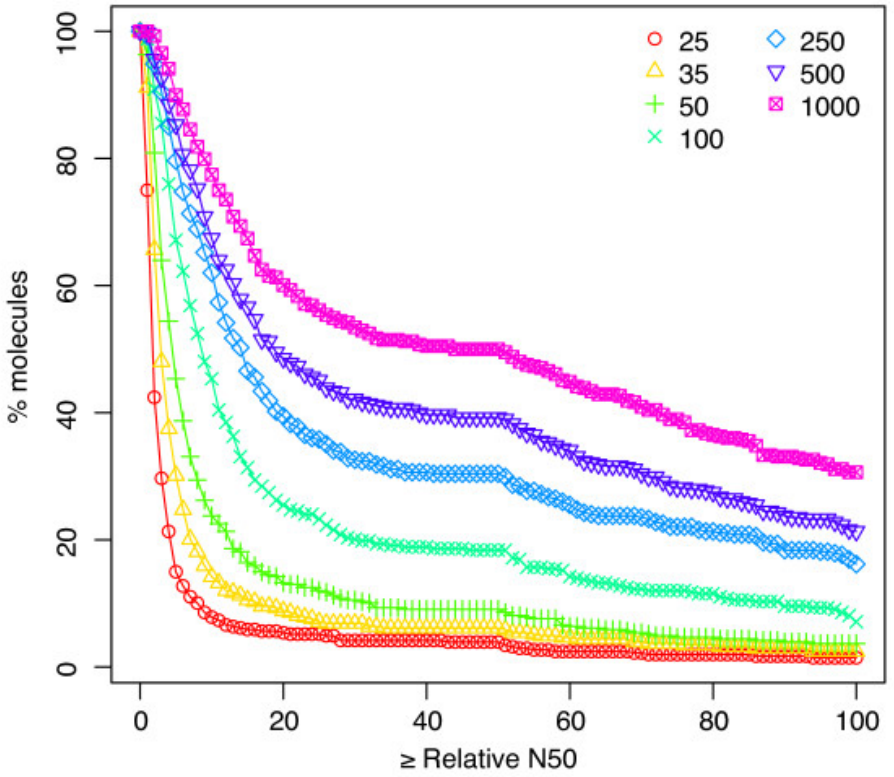
Genomes are not random k-mers

Genome uniqueness




E. coli genome k-mer analysis
4,443,317 (96%)
4,639,674
total positions
in genome

Absence of repeat sequence simplifies assembly



From 384 completely assembled bacterial and archaeal genomes

Relative N50 = N50 / chromosome length

Kingsford *et al*. *BMC Bioinformatics* 2010 **11**:21 doi:10.1186/1471-2105-11-21

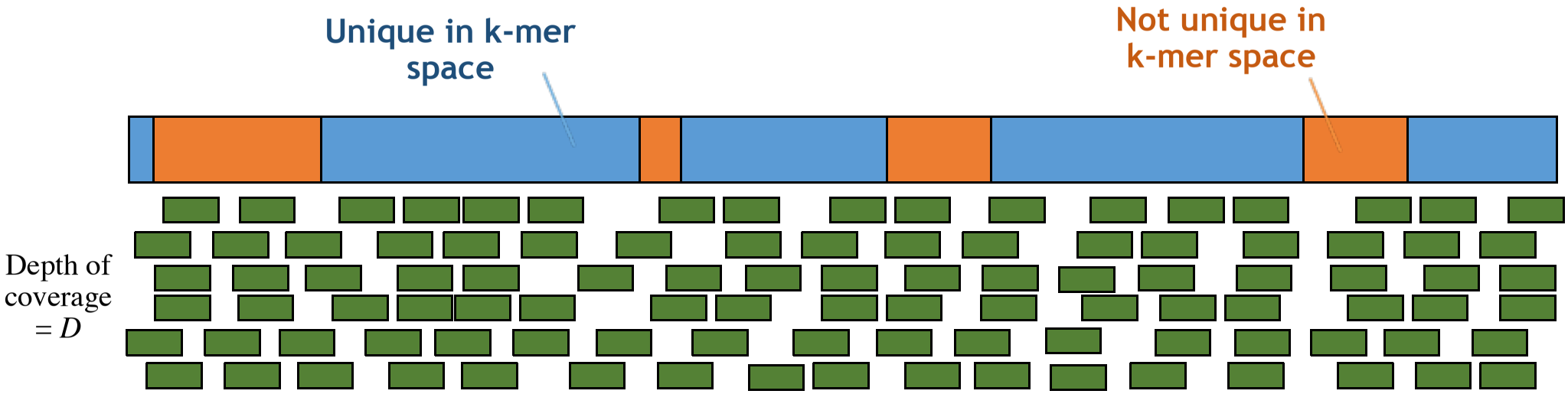
Table 1B. Fraction of *K*-mers having a unique placement on the genome

K	*E. coli* (%)	*S. cerevisiae* (%)	*A. thaliana* (%)	*H. sapiens* (%)
200	98.5	95.9	97.4	97.6
160	98.3	95.6	97.1	97.2
120	98.2	95.2	96.6	96.6
80	98.0	94.7	95.4	95.2
60	97.8	94.4	94.4	93.1
50	97.7	94.2	93.4	91.2
40	97.6	93.9	92.2	88.3
30	97.4	93.5	90.4	83.4
20	97.0	92.9	86.5	71.8
10	0.0	0.0	0.0	0.0

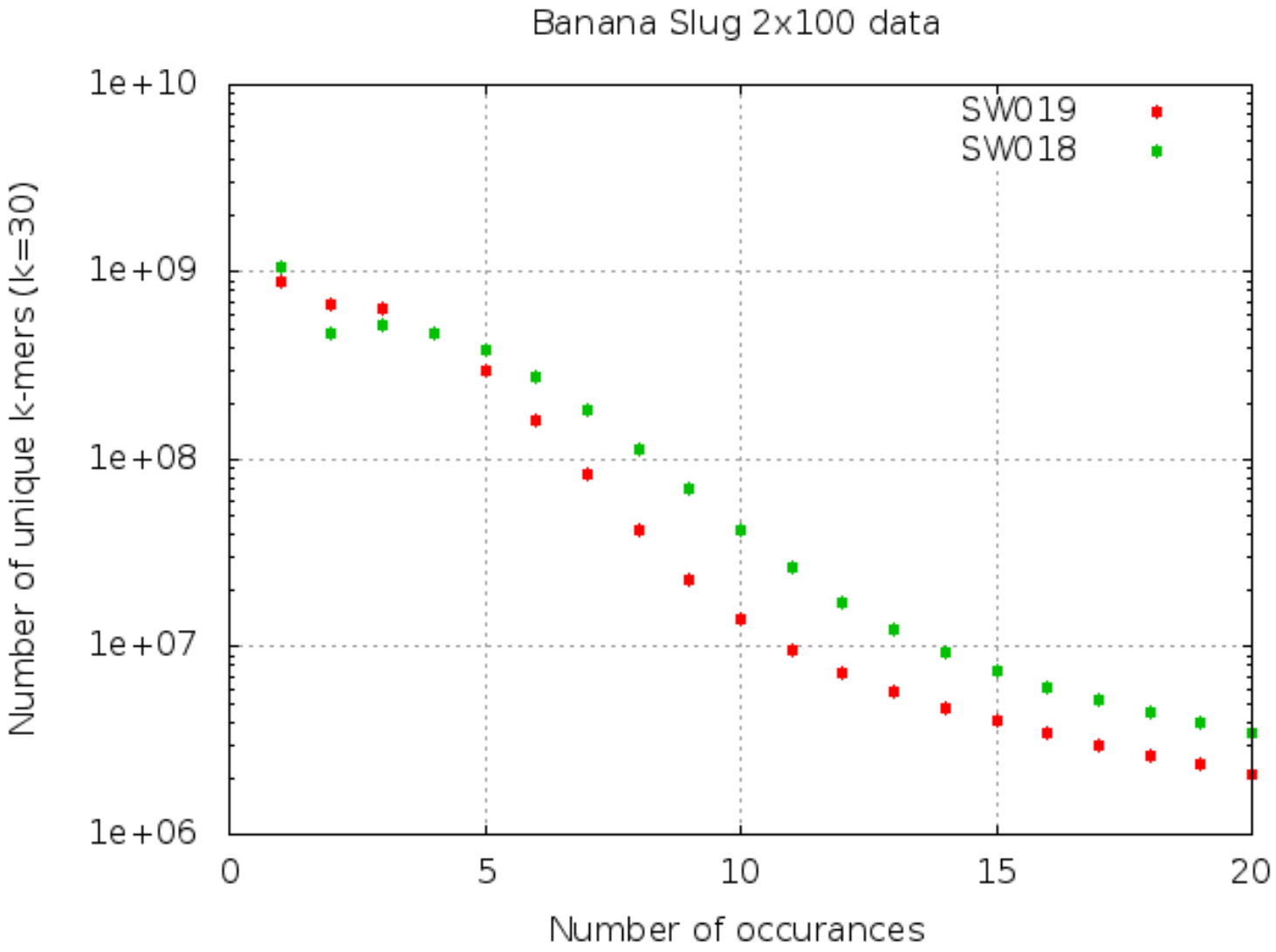
For a given *K* and a given genome, we show the fraction of its *K*-mers that have a unique placement on the genome. Values were estimated using a sample size of 10^4.

Butler et al, 2008 *Genome Research* **18:**810

K-mer spectra to estimate genome size



Total genome size =~ *S* / *D*
Where *S* = total sequence data &
D = mode of k-mer spectra


Banana Slug 2x100 data
SW019
SW018
1e+10
1e+09
1e+08
1e+07
1e+06
Number of unique k-mers (k=30)
0
5
10
15
20
Number of occurances

Other k-mer related questions about genome composition:

1. What fraction of the genome is in multi-copy repeats?
2. What fraction of the genome is heterozygous?
3. What fraction of the genome is in low-complexity regions?

K-mer spectra analysis for error correction

ACGTACGAGTCGATGATCGTCATGC

Low-quality base

What if this 25-mer occurred once in a dataset of 30-fold genome coverage?

What if a similar k-mer occurred 30 times in a dataset of 30-fold genome coverage?

Quake: http://genomebiology.com/2010/11/11/R116