"For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?""

-- George Box

H. influenzae genome is 1,830,138 bp

Base	Number	Frequency
A	567,623	0.3102
C	350,723	0.1916
G	347,436	0.1898
T	564,241	0.3083

Note that while we only counted bases on one strand, because of complementary we know the frequencies of the other strand.

GC content

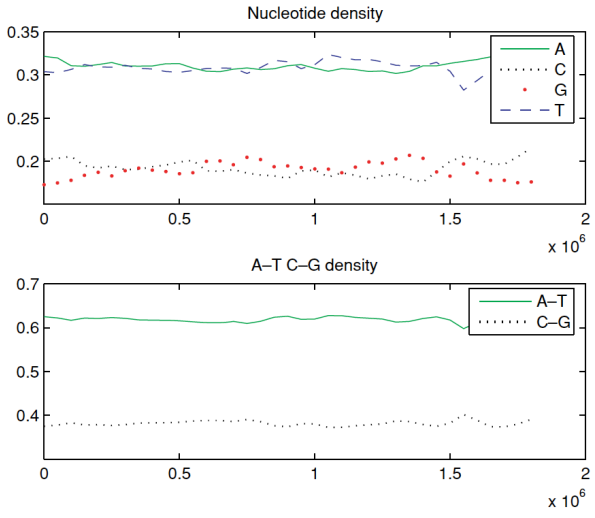
Because the content of G and C is typically similar, most genomes are reported as simply the GC content = p(G) + p(C).

Plus GC content isn't dependent on which strand you used to count p(G) and p(C).

Organism	GC content
H. influenzae	38.8
M. tuberculosis	65.8
S. enteritidis	49.5

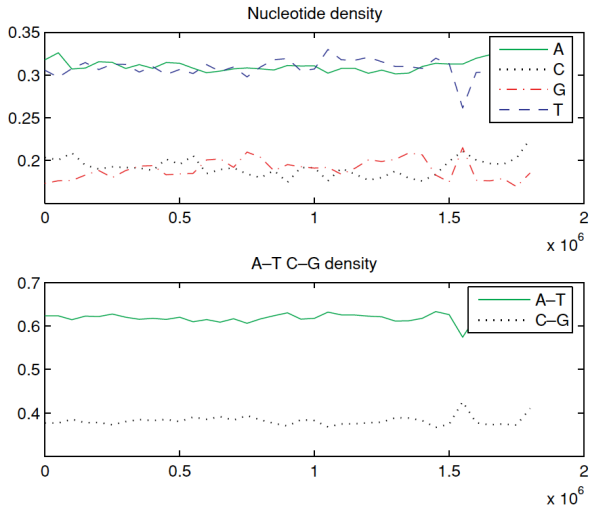
But what about local base composition?

k = 90,000

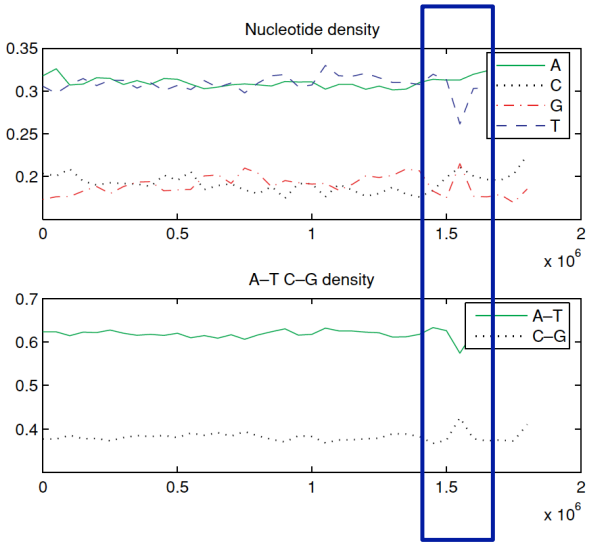


But what about local base composition?

k = 20,000



What do anomalies mean?



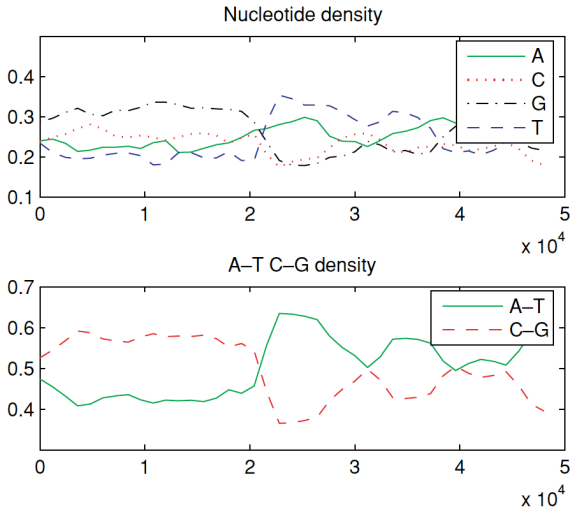
Methods to detect regions of statistical deviation from the average are referred to as *change point analysis*.

GC content

Because the content of G and C is typically similar, most genomes are reported as simply the GC content = p(G) + p(C)

Organism	GC content
H. influenzae	38.8
M. tuberculosis	65.8
S. enteritidis	49.5

An odd genome: lambda phage



So are the frequencies of nucleotides independent?

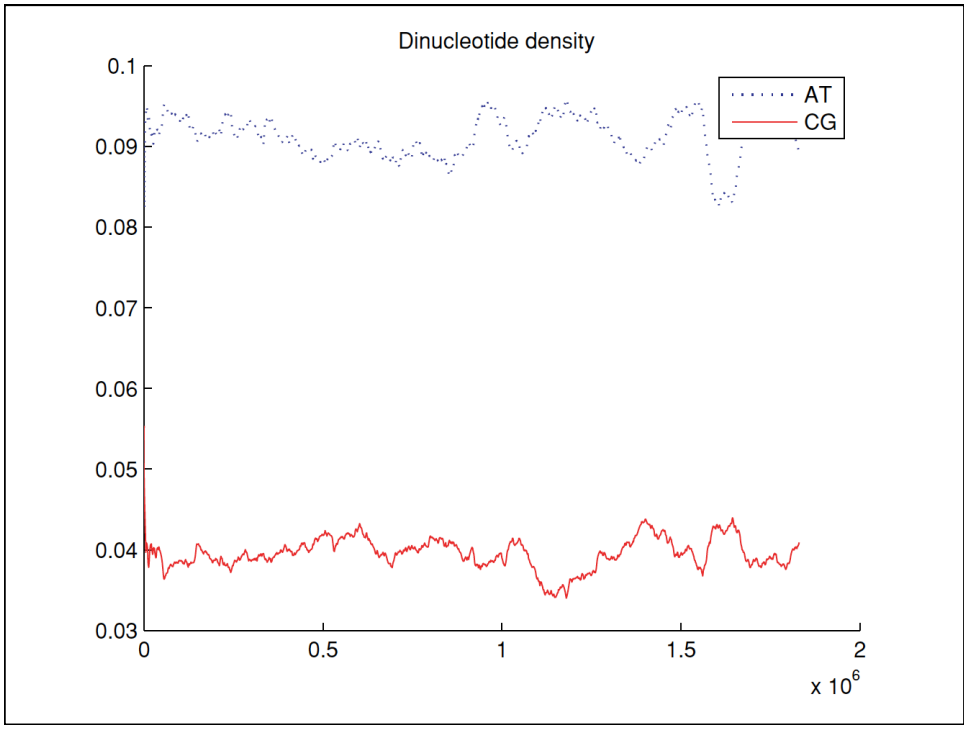
Consider the empirical frequency of all 2-mers:

	*A	*C	*G	*T
A*	0.1202	0.0505	0.0483	0.0912
C*	0.0665	0.0372	0.0396	0.0484
G*	0.0514	0.0522	0.0363	0.0499
T*	0.0721	0.0518	0.0656	0.1189

H. influenzae genome is 1,830,138 bp

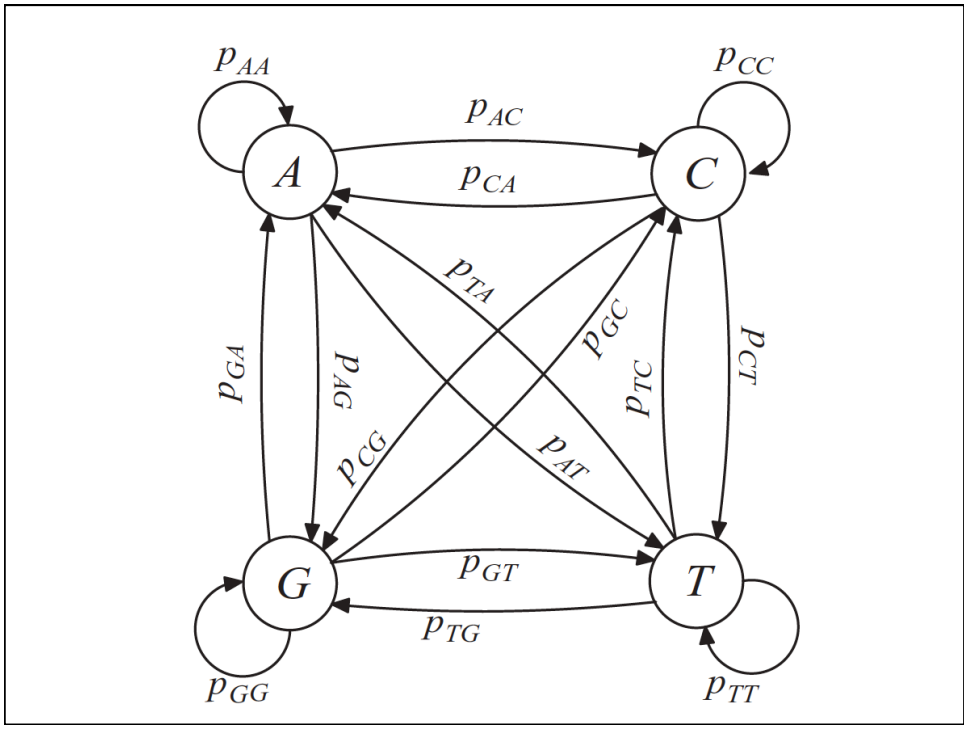
Base	Number	Frequency
A	567,623	0.3102
C	350,723	0.1916
G	347,436	0.1898
T	564,241	0.3083

Note that while we only counted bases on one strand, because of complementary we know the frequencies of the other strand.



Odds ratio of all 2-mers

	*A	*C	*G	*T
A*	1.2491	0.8496	0.8210	0.9535
C*	1.1182	1.0121	1.0894	0.8190
G*	0.8736	1.4349	1.0076	0.8526
T*	0.7541	0.8763	1.1204	1.2505


p_{AA}
p_{CC}
p_{AC}
A
C
p_{CA}
p_{TA}
p_{GC}
p_{GA}
p_{AG}
p_{TC}
p_{CT}
p_{CG}
p_{AT}
G
T
p_{GT}
p_{TG}
p_{GG}
p_{TT}