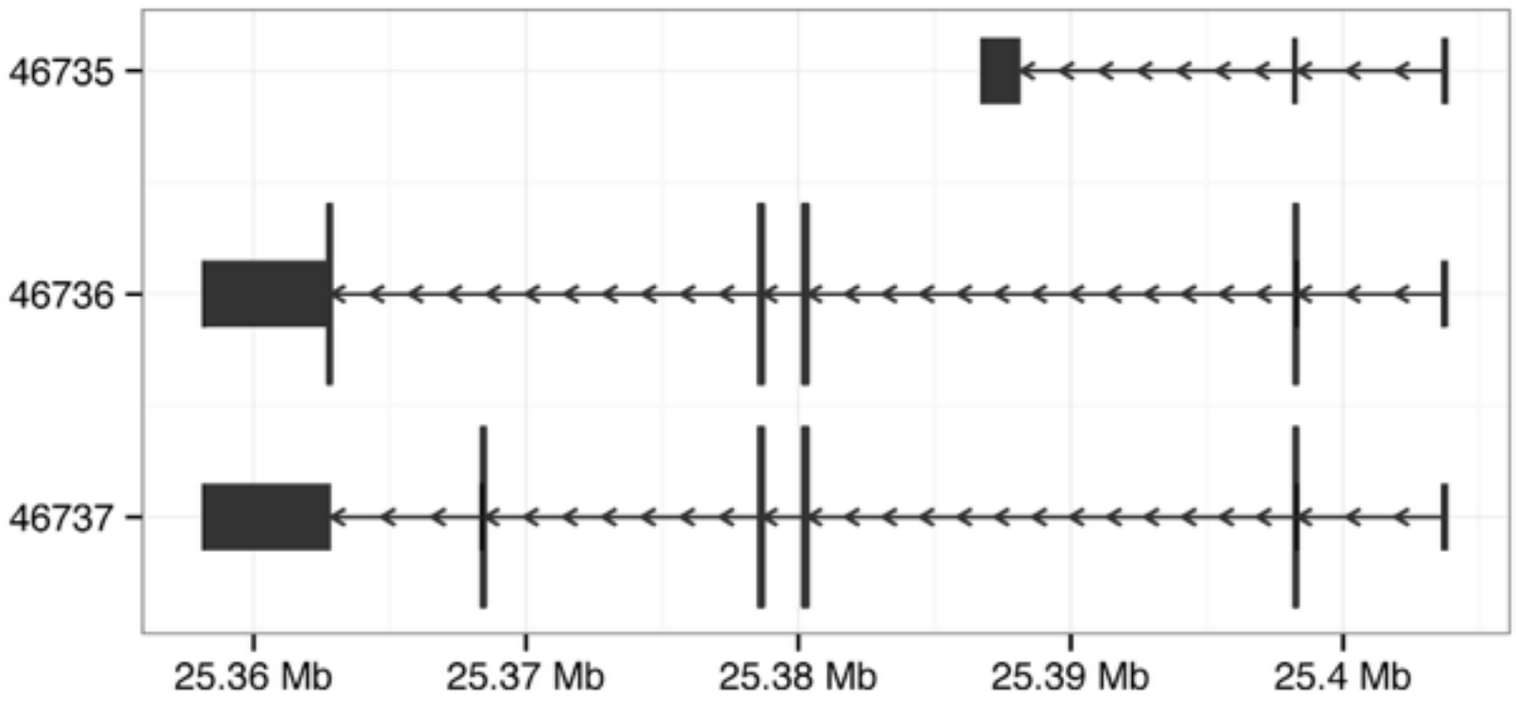
Genome Architecture

CMSC 702
Spring 2014

Genome is organized

- Remember the basic unit of study is the *gene*
  - These correspond to specific *regions* in chromosomes (genome)
  - These regions are further organized
    - exons/introns
    - transcripts
    - promoter regions

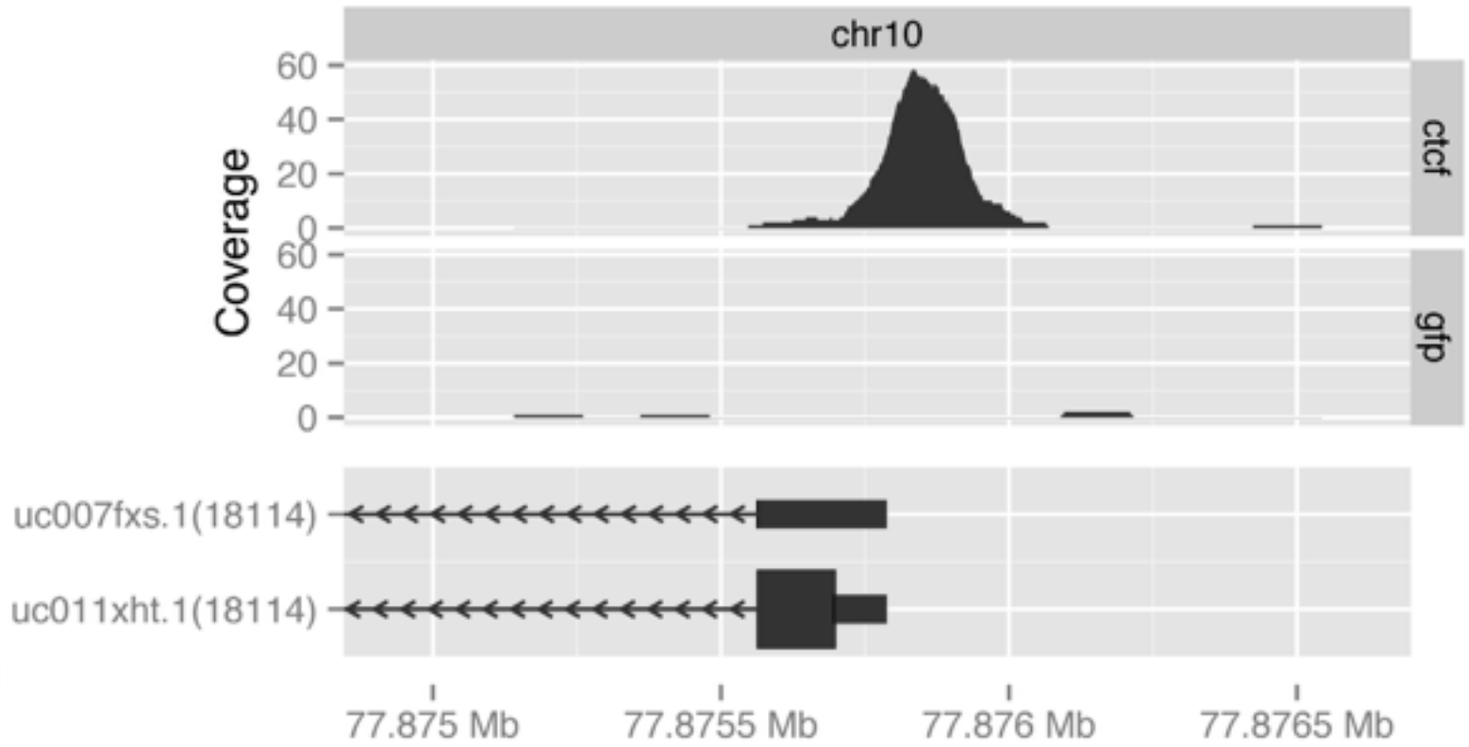


And more…

- Repetitive regions
  - (Alu, LINE, SINE elements)
- Telomeres, centromeres, …
- CpG Islands (more later)

Functional Genomics

- Many quantitative experiments are analyzed with respect to known genomic regions
  - RNA-seq: gene expression
  - ChIP-seq: transcription factor binding
  - DNAm-seq: measure methylation with sequencing

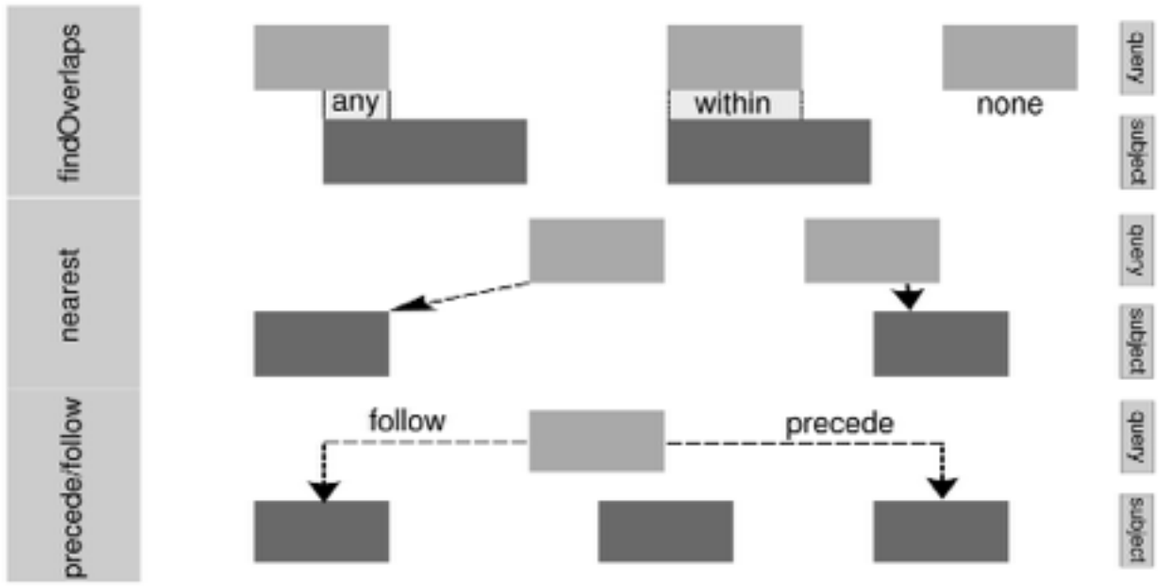


Regions as intervals

- Fundamentally these regions are represented as intervals in the genome
    - *GenomicRanges* in Bioconductor
    - *bedtools*
        - And combined into lists of intervals for more complex regions
            - (e.g., gene models)
- Common task in genomics is comparison, annotation and discovery of intervals

Borrows from spatial analysis

- Common region operations
  - extend, reduce
- A really important one is *overlap*
- Interval trees are fundamental data structures in use



R/Bioconductor

- Provides the *IRanges* and *GenomicRanges* packages to support this
- *IRanges:* lists of intervals
- *GenomicRanges:* lists of intervals in genomes (divided into chromosomes)
- *SummarizedExperiments:* Quantitative data encapsulation. Features are *GenomicRanges*

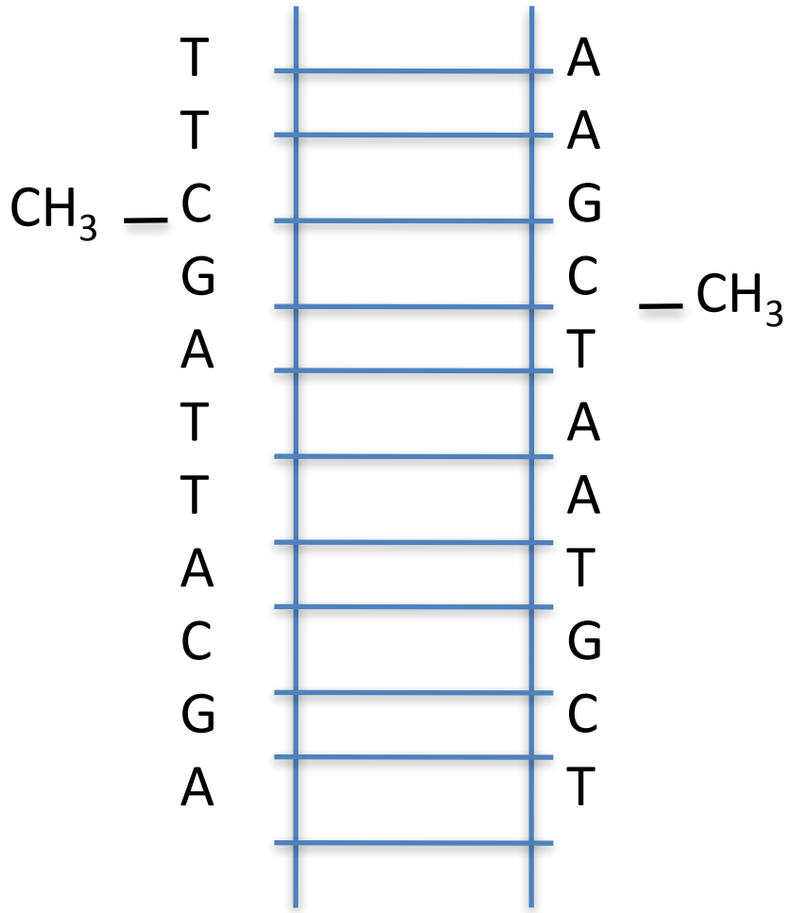
The future

- Lots of new data and findings made
  - Expression, Methylation, Conformation.
  - In many tissues, many conditions.
- All of these are organized around regions
- The future: searching the genome for regions of interest:
  - systems, stats/ML, algorithmics

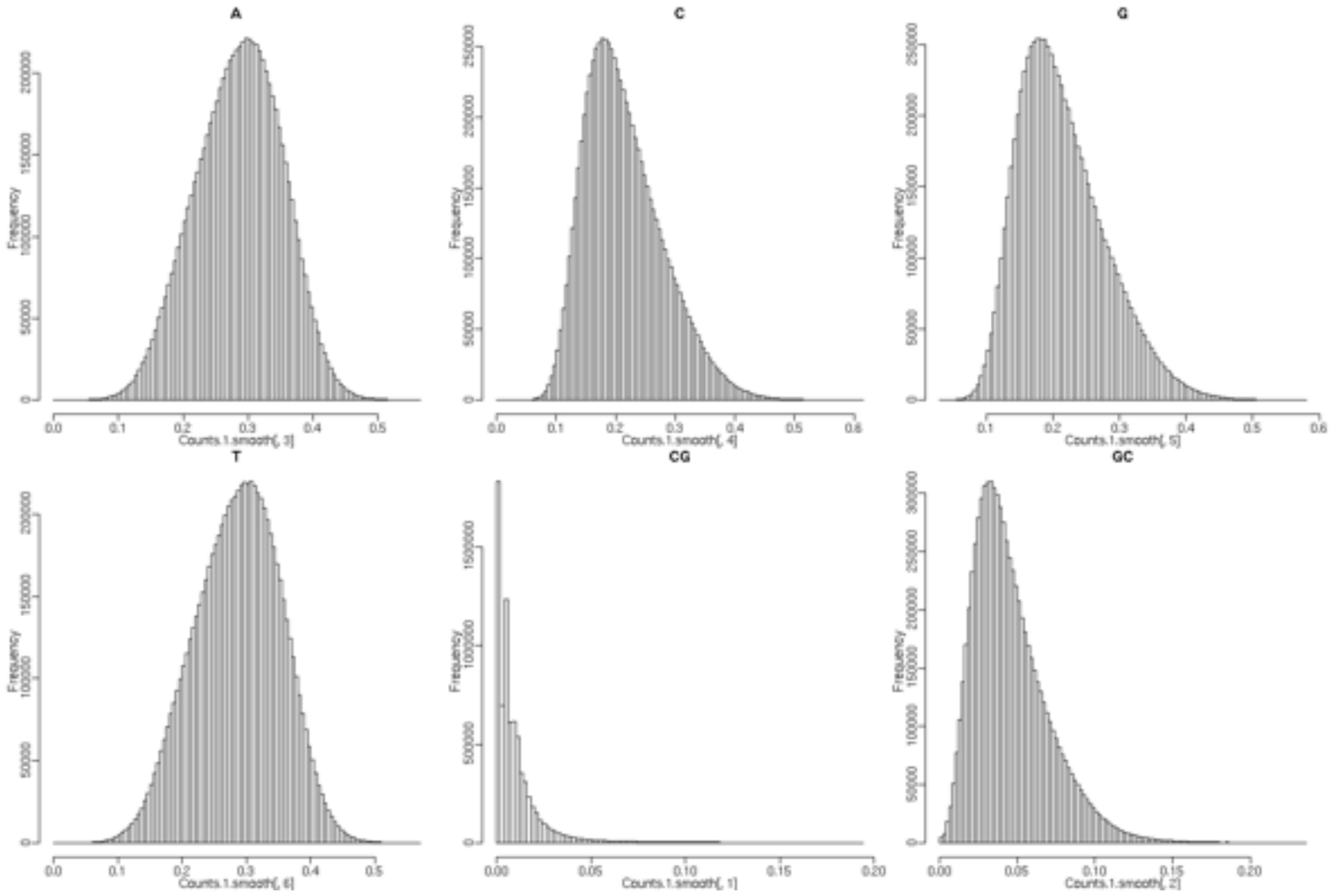
One Example:
CpG Islands

How are regions defined?

Methylation



CpGs are depleted

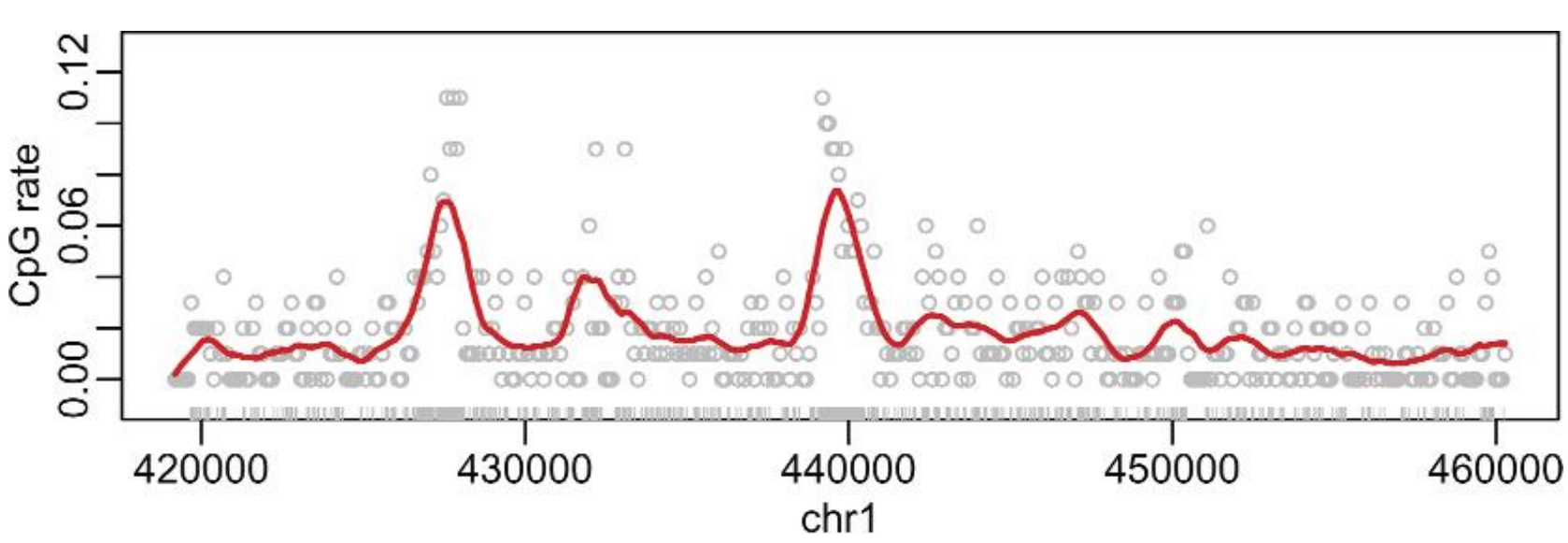


One basic way of characterizing regions is by their *sequence* content: the *proportion* of A,C,G or T in a given region, or the proportion of *di*-nucleotides, e.g., *CG.*

CpGs are depleted and they cluster

CpGs are depleted

Remaining CpGs cluster into *islands* enriched near promoters

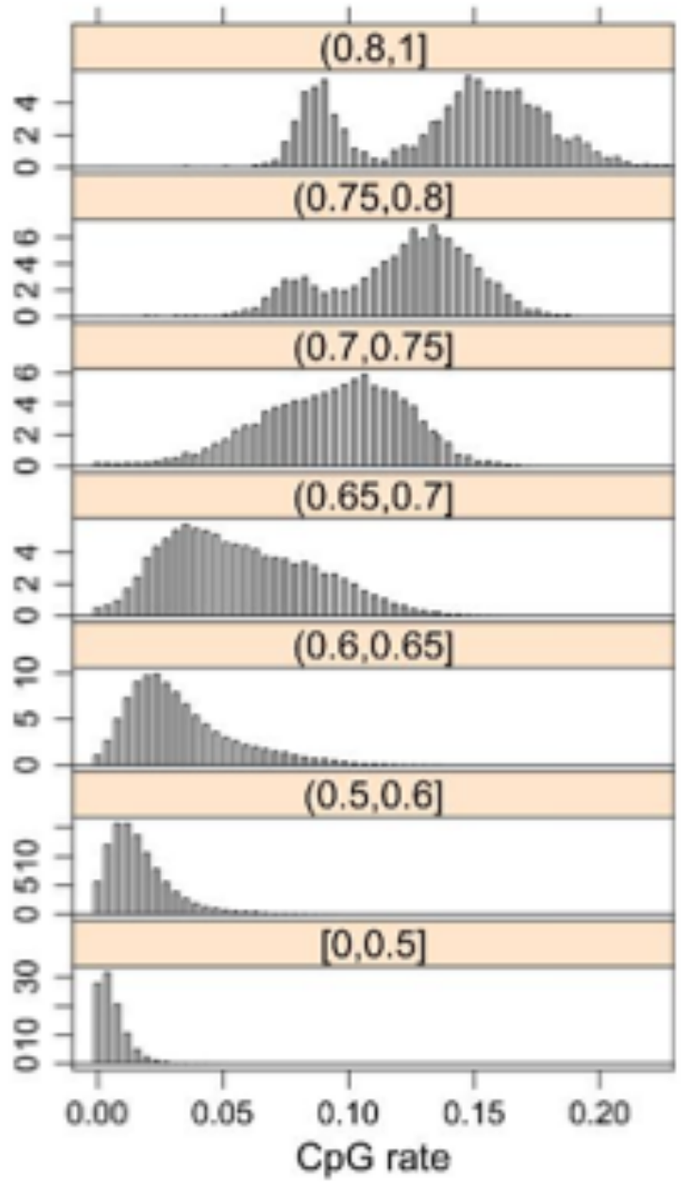


Remaining ones cluster

Two modes:

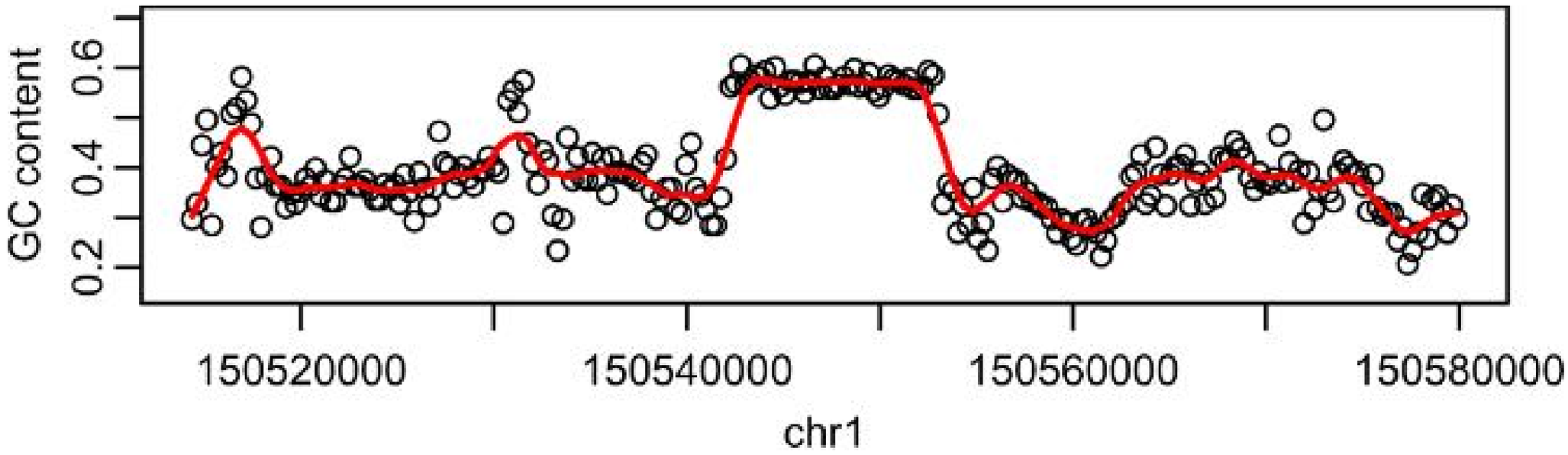
high CpG rate

low CpG rate



Proportion of CpG's stratified by CG content.

GC content *also* clusters



CpG Island definition

Gardiner-Garden and Frommer

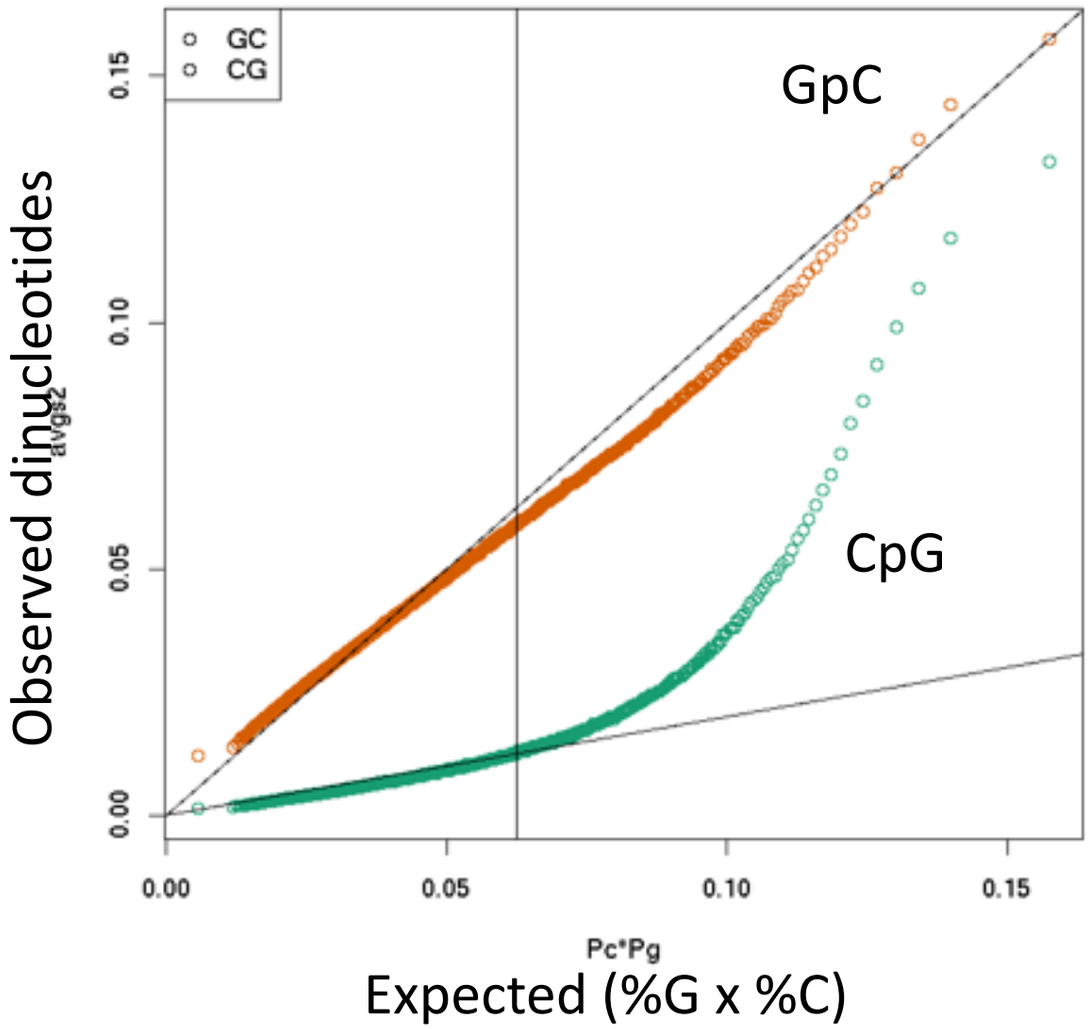
- N > 200
- GC-content > 50%
- obs/exp > 0.6
- Lists contain 20,000 CGI

Takai and Jones PNAS 2002 use a stricter definition

HMM based definition

- Irizarry et al. (2009) *Mammalian Genome*
- *Wu et al (2010) Biostatistics*
- *Lists contain 100,000 CGI*

Observed versus expected



Expected (%G x %C)

Modeling via HMM

- See lecture notes