

Are Gene Locations Clustered on Chromosomes?

Naomi S. Altman^{*}

Eli Walters^{*}

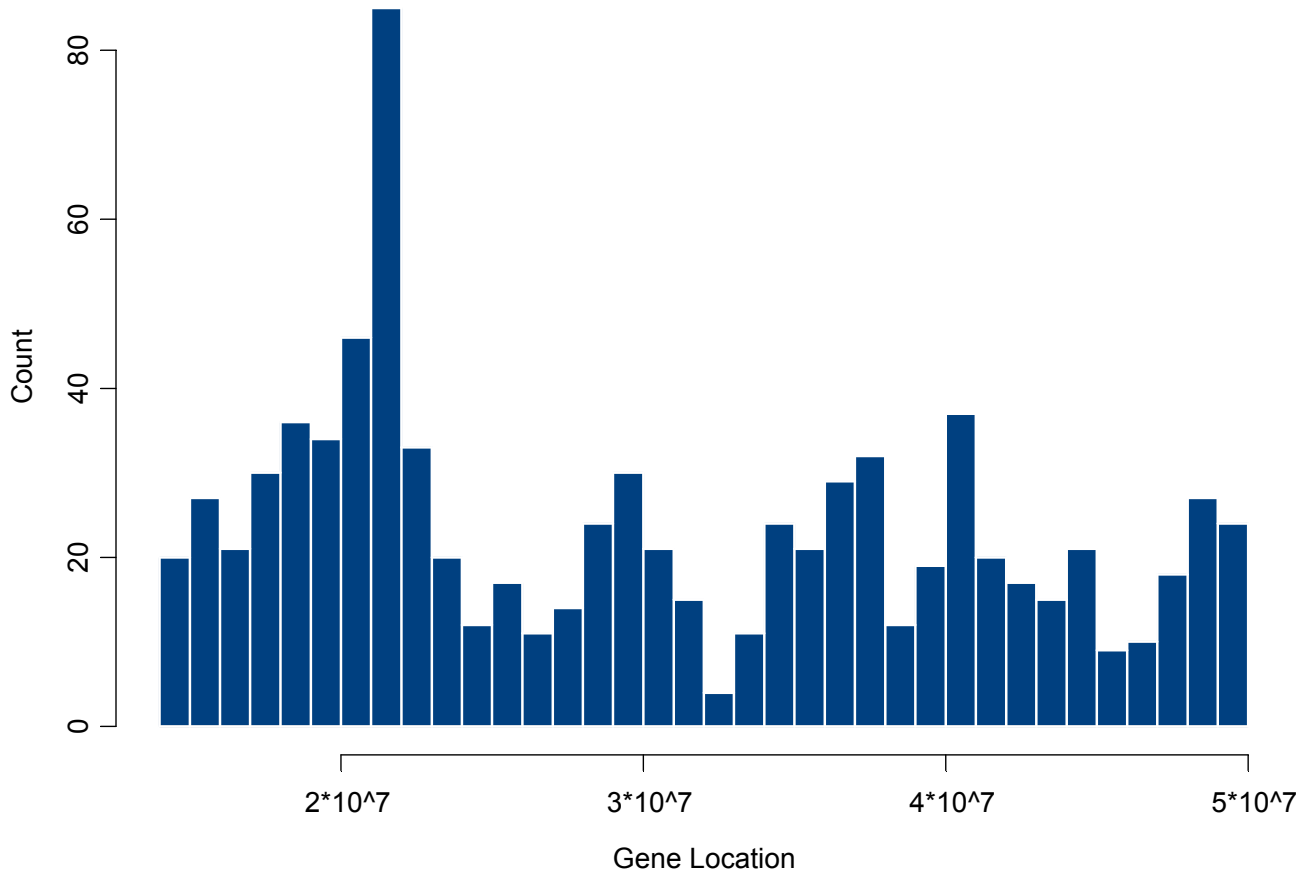
Laura Elnitski[#]

Pennsylvania State University

*** Statistics**

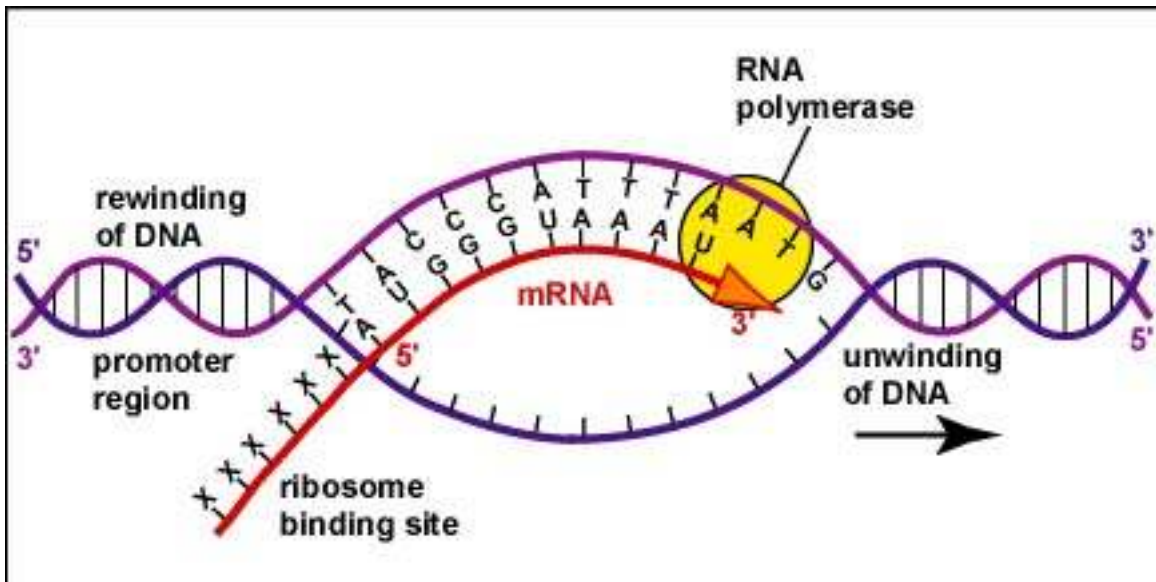
Biology

Genes appear to cluster on chromosomes



789 Gene Start Locations on Human Chromosome 22 – Hu 22

Genes Seldom Overlap

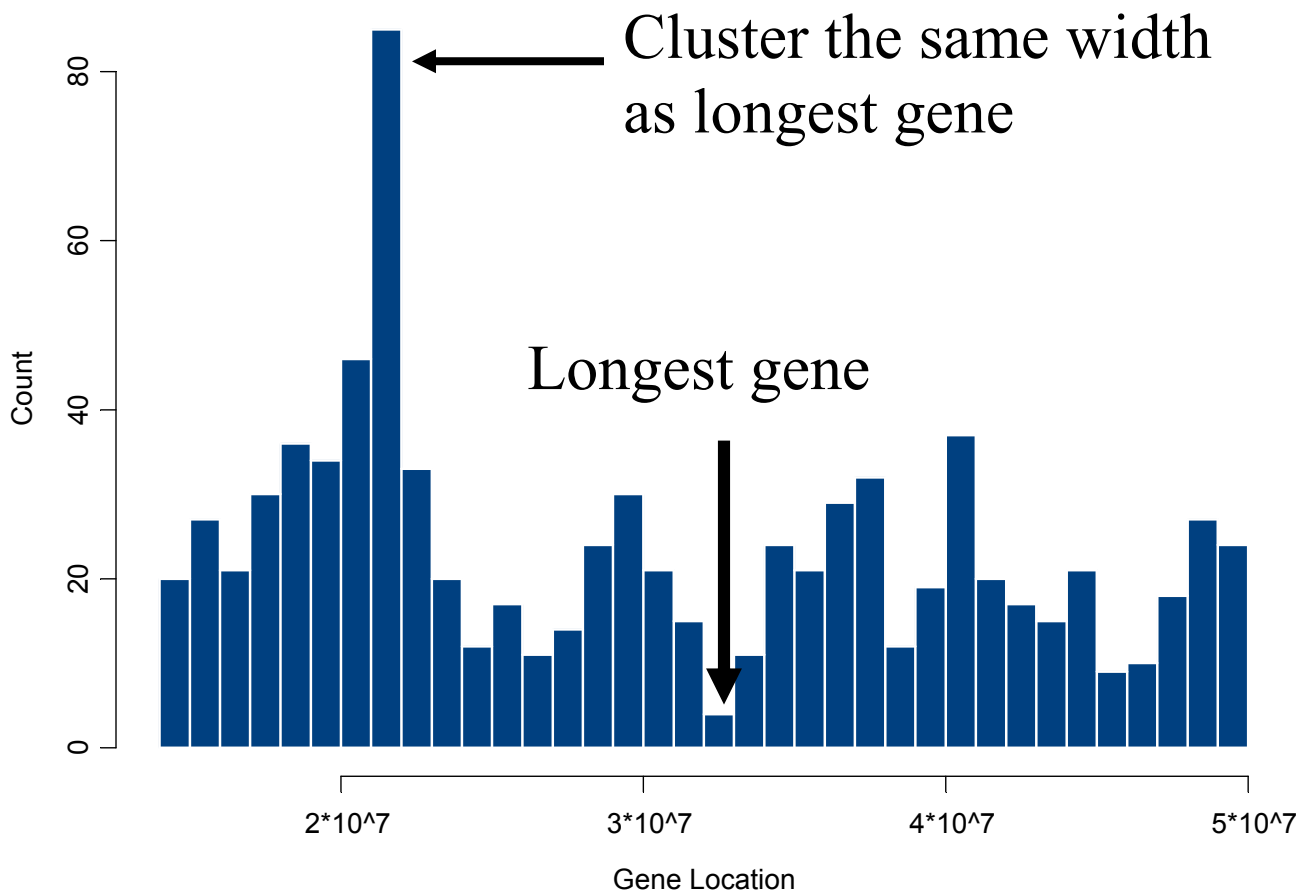


Each gene lies on one strand of the DNA double helix. Genes on the same strand almost never overlap. Most genes are on the same strand.

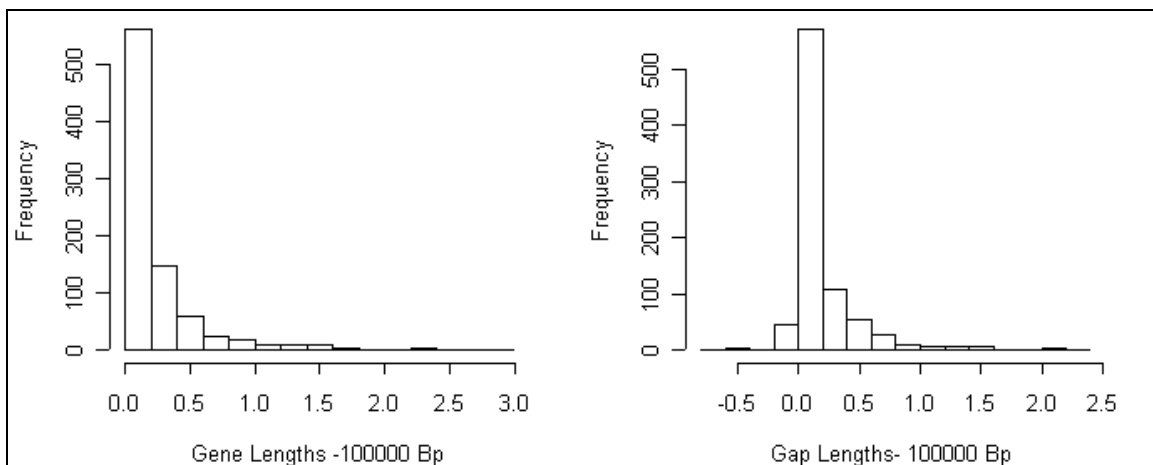
**On Hu22: Shortest gene: 37 bp
Longest gene: 647338 bp**

A lot of short genes could sit in the same space as the longest gene!!

We need to consider gene length when we look at clustering

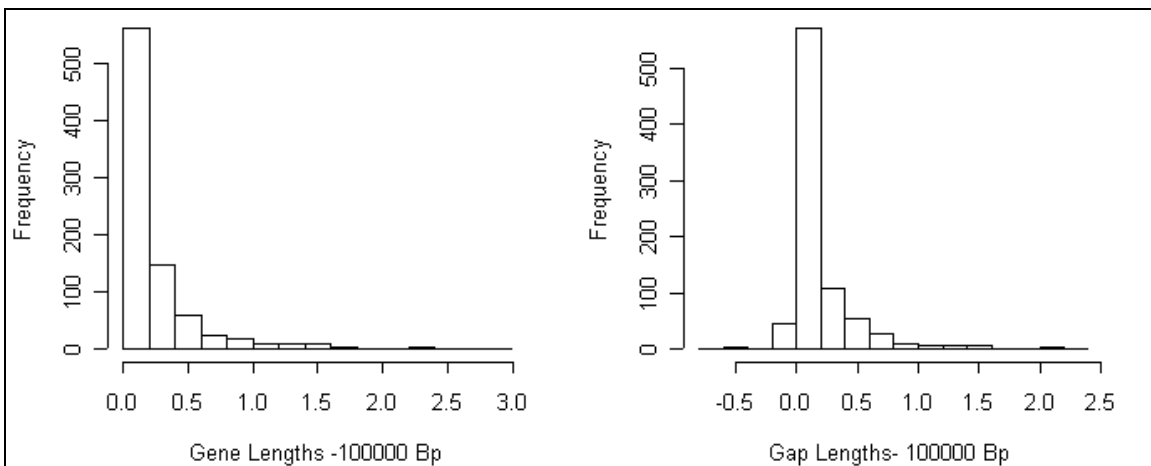
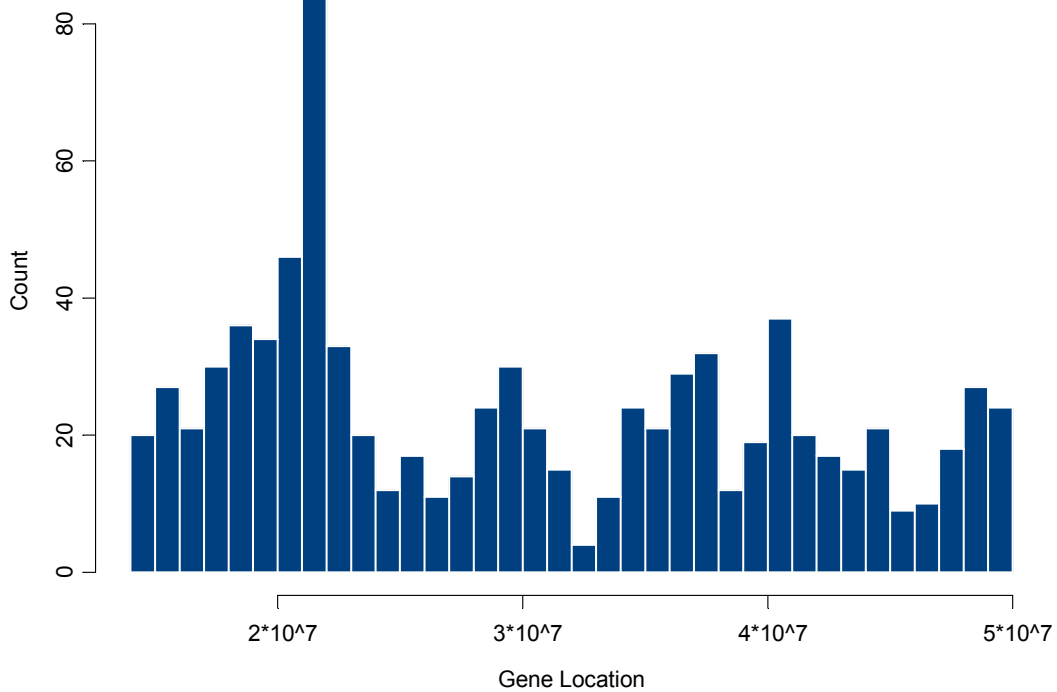


Genes and the Gaps Between Them Have Lengths

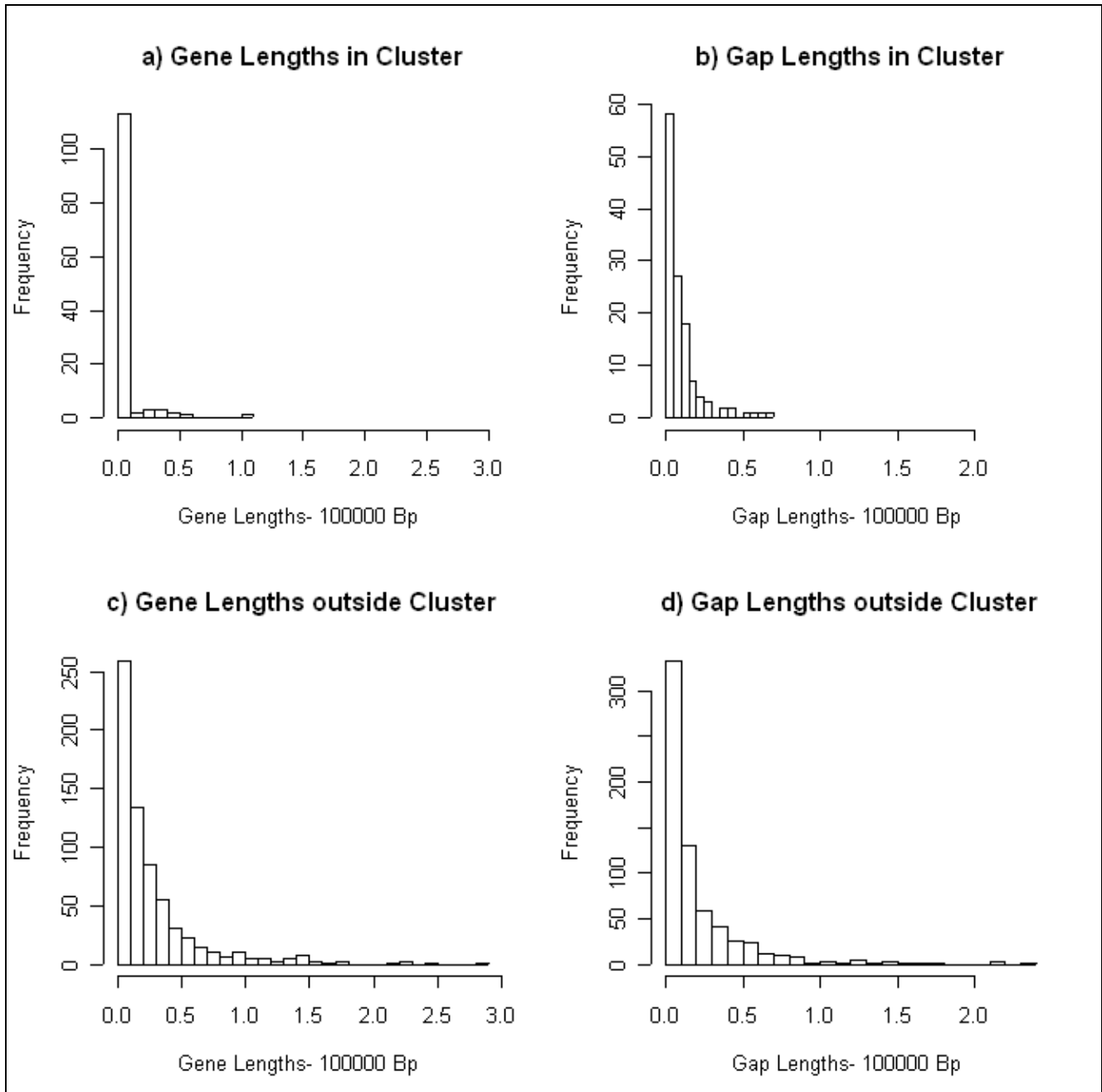


**Histograms of Gene Length and Gap
Length for genes on Hu22**

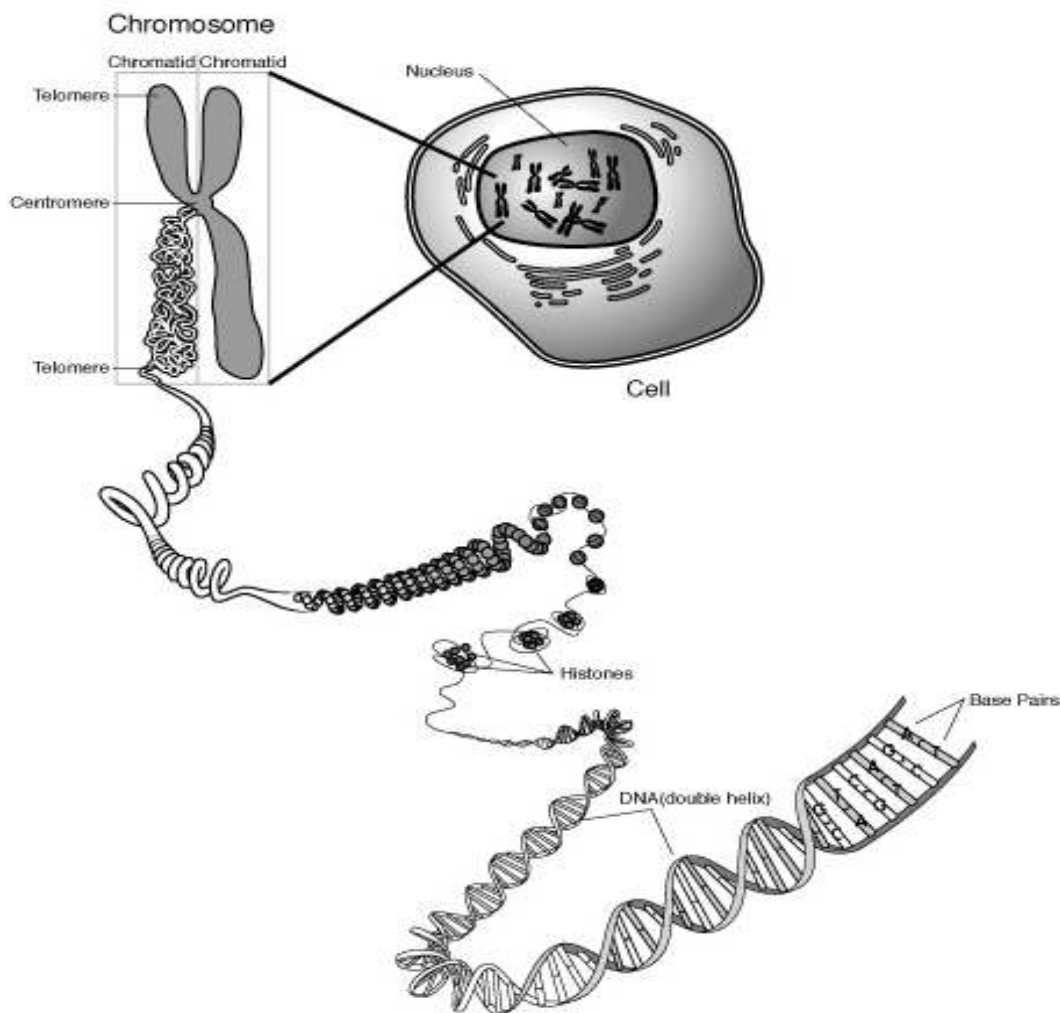
Short Genes or Clustered Locations?



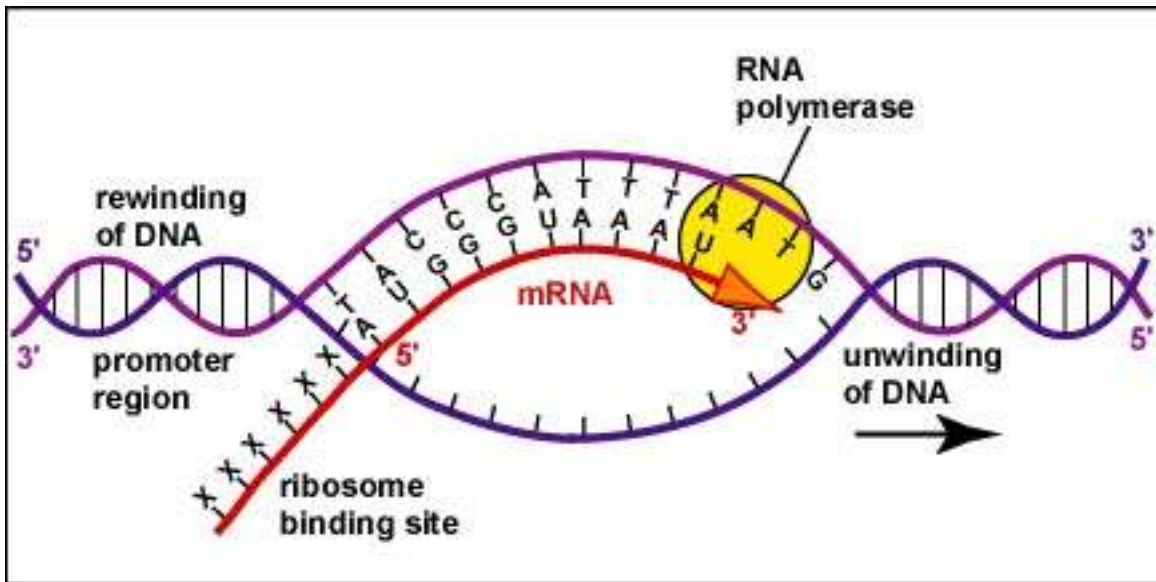
The Genes in the Dominant Cluster are Short and have Short Gaps



Why do we care about gene location clustering?



Genes can interact through spatial organization including proximity due to adjacency and due to folding



**and when a gene expresses,
transcription factors “unzip” the
chromosome and initiate
transcription. Adjacency might
promote or inhibit co-expression**

**So the organization of genes
along the chromosome has
biological implications**

Tests of Clustering

Tests of Uniformity

Our approach:

If we reject uniform distribution of gene locations we accept clustering.

There are lots of tests of uniformity.

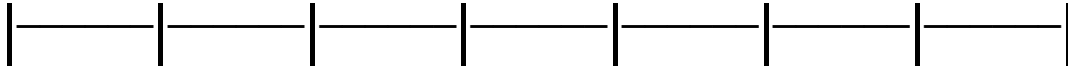
We select 3:

Chi-squared test

Kolmogorov test

Cramer – von Mises test

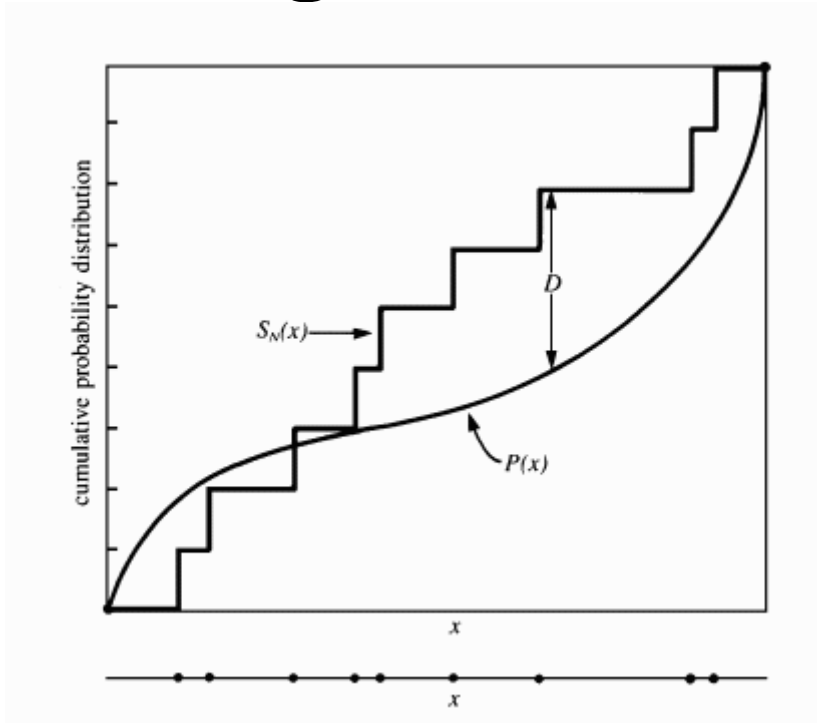
Chi-squared test



Distance along chromosome

Cut the chromosome into K equal length “bins”. Count the number of genes in each bin. Consider the differences between the counts and the average.

Kolmogorov and CvM tests



Distance along chromosome

“Staircase” is the percentage of genes before this distance. Curve is the percentage under a hypothesized distribution (which in our case is a “staircase” with equal sized steps).

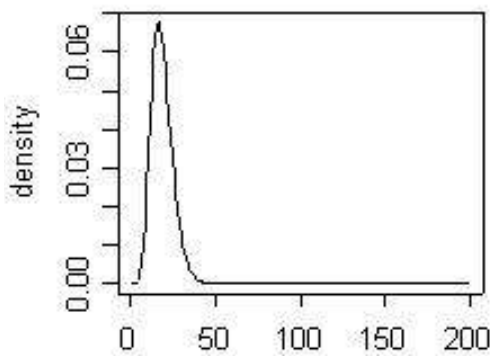
The Kolmogorov test is the maximum difference. The CvM test is the area between the curves.

Null Distribution

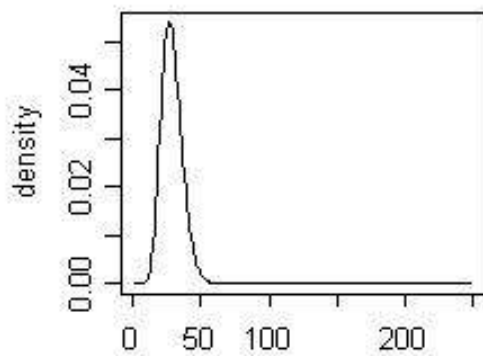
Tabulated null distribution assumes that gene locations are independent.

Gene locations are not independent due to the overlap problem and varying lengths of the genes.

Comparison of Tabulated Null Distribution and Simulated Null from Data Resembling Hu22



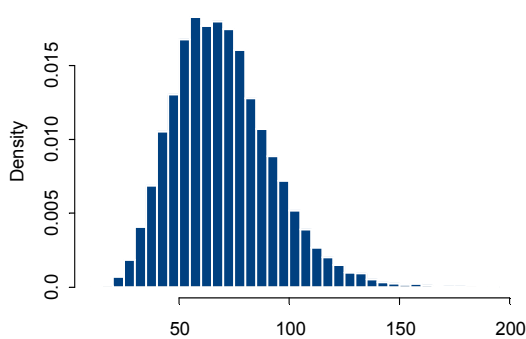
Chi-squared Distribution on 19 d.f.



Chi-squared Distribution on 29 d.f.

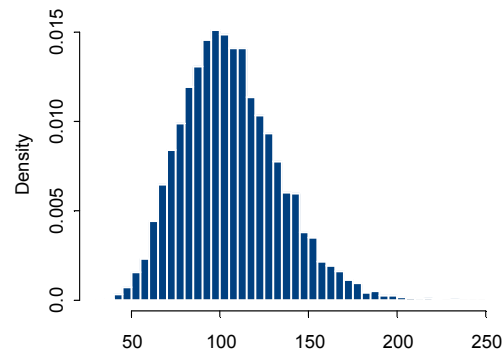
Null distribution under independence

(a) Chi-Square with 20 Bins



Chi-Square with 20 Bins

(b) Chi-Square with 30 Bins



Chi-Square with 30 Bins

Null distribution under observed gene and gap length distribution

Power of the Tests

We considered:

Cluster location (center or edge)

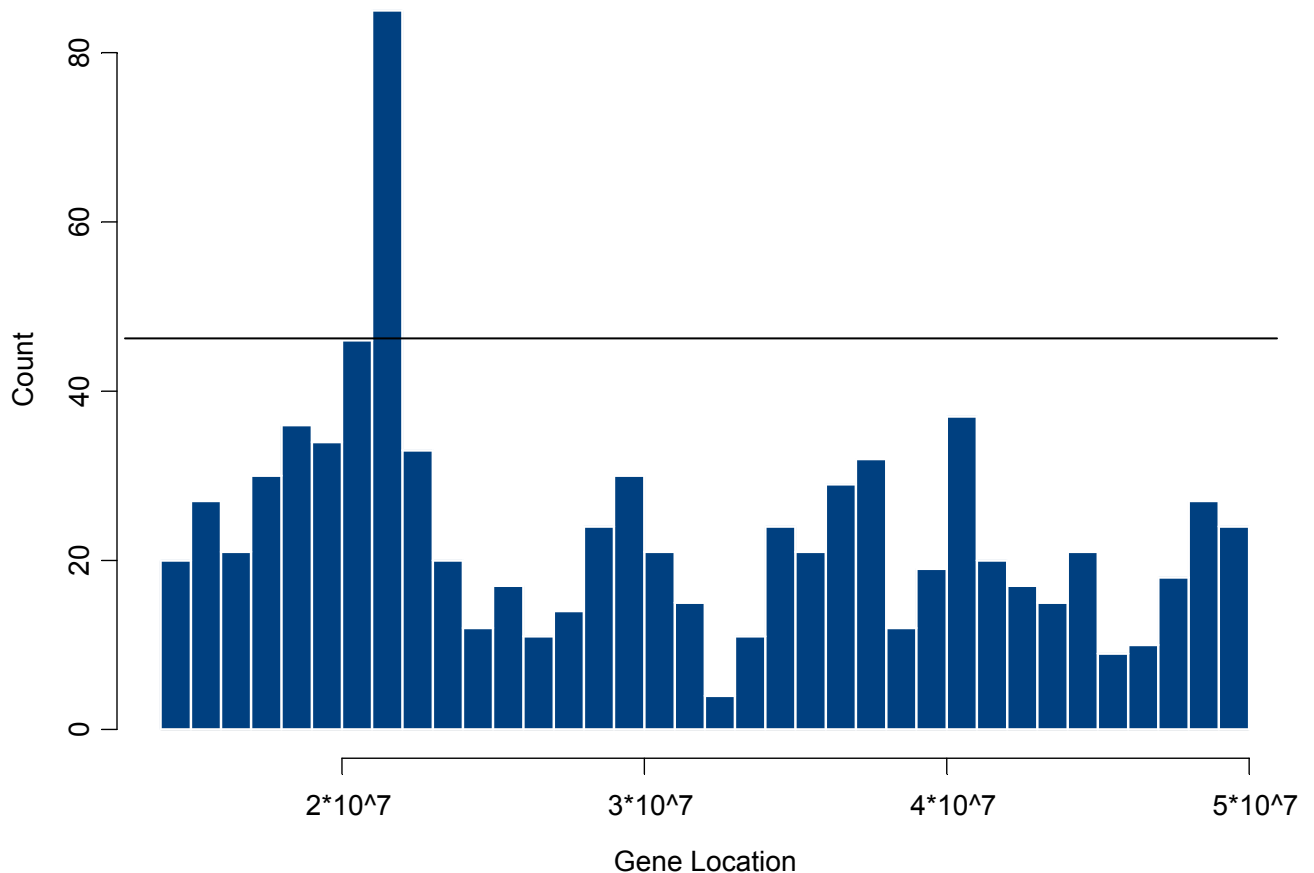
Percentage of genes in the cluster

Gene and Gap length distributions

The winner:

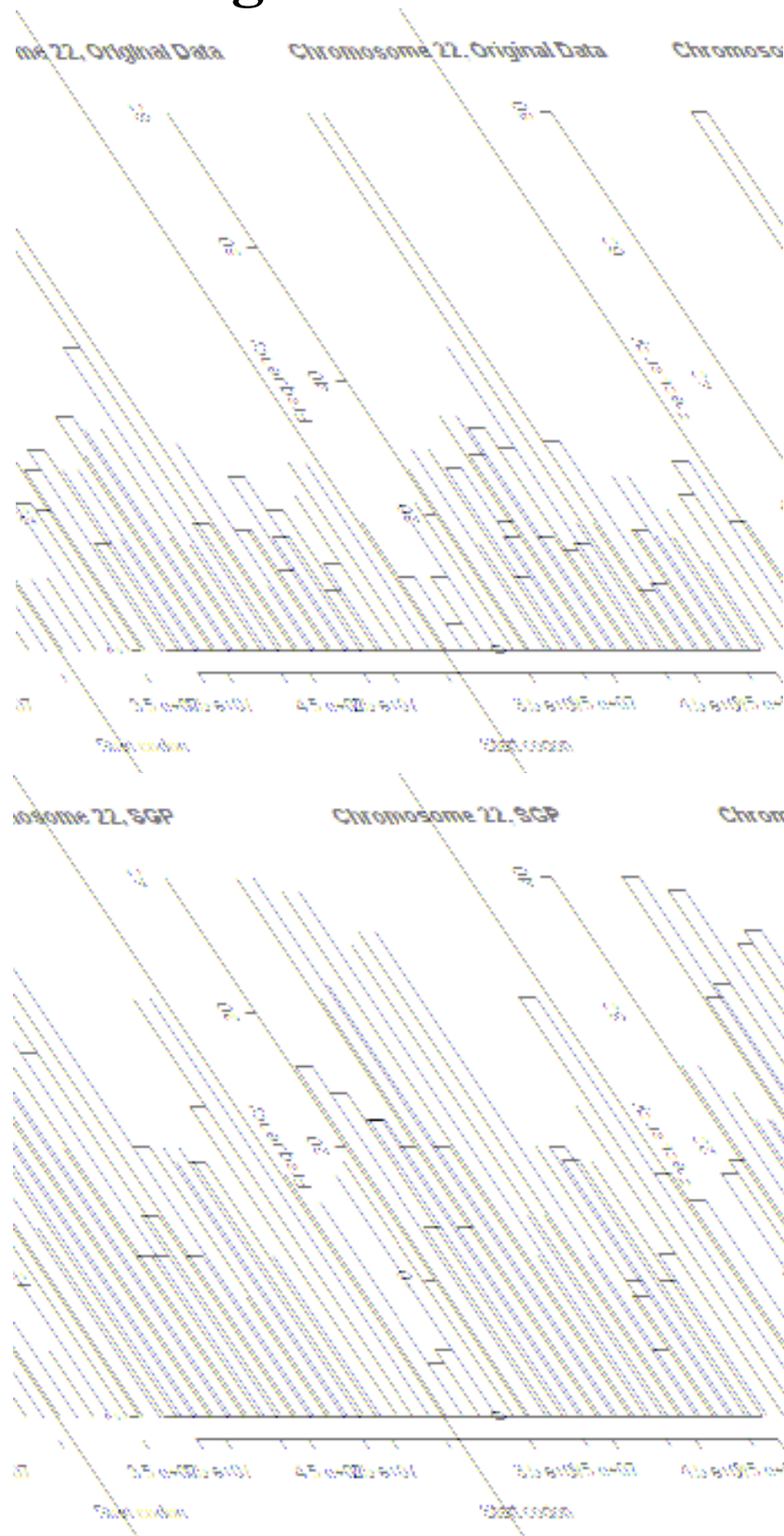
Chi-squared Test

Are genes clustered on Hu22?



All the test statistics “find” the obvious cluster. The horizontal line is the bootstrap 95th percentile of the “maximum bin count”

What is the biological significance of gene clusters?



Conclusions

- **Genes appear to cluster on human chromosomes**
- **Simulate null gene distribution by sampling gene and gap lengths**
- **Chi-squared test works well (with bootstrap null)**
- **Chi-squared test can also be used to determine location of multiple compact clusters using cell z-scores $(O_i - E_i) / \sqrt{E_i}$.**
- **Tests based on cdf do best when cluster occurs near the ends of the chromosome**
- **Even for “well-annotated” chromosomes, the data are very much in flux**

Acknowledgements

Pallavi Eswara (biology, PSU) prepared the original data using the GALA routine, Giardine et al., 2003