Revisiting Gene Deserts: New Definitions and Investigations of Overlapping Features

The Jackson Laboratory

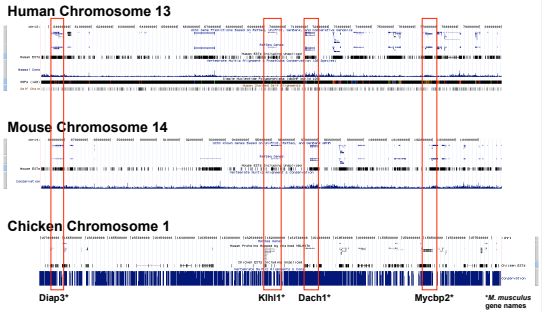
Nicole P. Leahy, Alex Ellison[1], Cheryl Zapata[2], Joel H. Graber

Center for Genome Dynamics, The Jackson Laboratory , Bar Harbor, Maine 04609 USA

Introduction

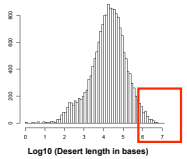
- Protein coding genes in metazoan genomes are not uniformly distributed
- Existing studies of "gene deserts" (large regions separating protein coding genes):
  - used a non-robust and somewhat arbitrary definition for deserts..
  - found that deserts, and the genes that flank them are frequently evolutionarily conserved.
  - found that the genes flanking deserts were frequently "developmentally important," and that these genes also frequently had highly conserved UTR sequences, possibly indicating conserved post-transcriptional regulation
- Conventional wisdom is that deserts hold critical regulatory sequence elements that control their flanking genes.

Deserts and Flanking (or Embedded) Genes are Evolutionarily Conserved



Existing Classifications of Gene Deserts are Arbitrary and Volatile with Genome Annotation

- The current standard.definition of a gene desert is contiguous genomic sequence with no protein coding genes with minimum length 500k-640k
- This definition has two principal problems:
  - It is arbitrary, with threshold set by total summed desert size



  - It is not robust with changing genomic annotations: discovery of gene size can abrogate existing deserts

Computational Methods

Dynamic Programming to Identify Gene Deserts

- We dynamic programming based on the complement of the gene density as the kernel, offset by a base density to provide the average negative value required for a stable dynamic programming solution

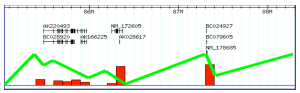
$$S_i = \alpha_0 - \rho_i$$

S_i = dynamic programming kernel
α_0 = base density
ρ_i = density in the i^{th} genomic window

- Gene density is defined in windows, currently either 10,000 or 100,000 bases, based on either (all drawn from the UCSC browser "known genes" table):
  - Exon density
  - Coding-only exon density
  - Transcribed base density, including all exons and introns
  - Gene count
  - Transcription start site count
- The dynamic programming is done in a Smith-Waterman formalism, searching for local best desert blocks:

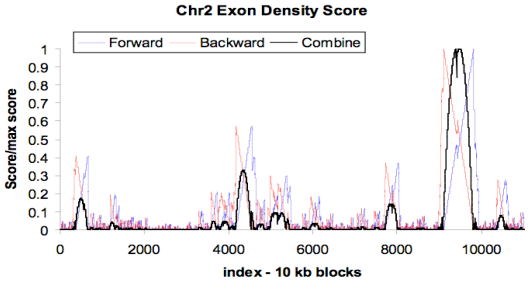
$$F_i = \max\begin{cases} 0 \\ F_{i-1} + S_i \end{cases}$$

F_i = desert score for the i^{th} genomic window
F_0 = 0



Red bars: windowed density
Green line: DP solution

- The dynamic programming is run forward and backward and the product Forward*Backward gives the final desert score at each window

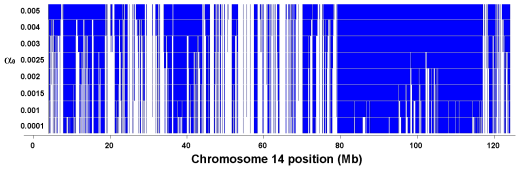


New Desert Definition Consequence: "Gene Oases"

- Our updated desert definition results in genes (or clusters of genes) that are embedded within deserts, or "gene oases"
- Gene oases provide new sets of genes that can be analyzed as groups to better delineate the functionality that is controlled by

Results

Desert size is a function of analysis parameters

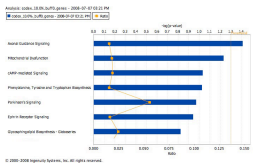


Gene oases have biased gene content

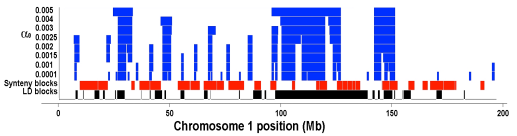
Th α_0 = 0.025, desert sum = 10% of genome, all embedded genes

VLAD (hypergeometric)

Ingenuity Pathways Analysis



Preliminary studies indicate overlap between gene deserts and other block-defining features



Conclusions and Acknowledgements

- Our dynamic programming approach to gene desert identification is robust with changes in genome annotation
- The updated desert definitions include embedded "gene oases"
- Genes in oases are biased towards a number of developmentally important functions, *e.g.*, cell signaling, axon guidance and formation, and cell-to-cell adhesion
- Funding for this project comes from NIGMS Center grant P50 GM076468

[1] Connecticut College, New London, CT 06320
[2] North Carolina State University, , Raleigh, NC 27695