

# **Genomic Dark Matter...**

## **The Dark Side of Life**

Abiy Dejenee  
Patrick Phepa  
Simon Johnstone-Robertson

# Overview

- What is Dark Matter?
- Plant Dark Matter (Simon)
- CNEs (Abiy)
- Heterochromatin (Patrick)



# Overview

- What is Dark Matter?
  - How much is there?
  - Putting it into perspective
- Plant Dark Matter (Simon)
- CNEs (Abiy)
- Heterochromatin (Patrick)



# Overview

- What is Dark Matter?
- Plant Dark Matter (Simon)
- CNEs (Abiy)
- Heterochromatin (Patrick)



# Overview

- What is Dark Matter?
- Plant Dark Matter (Simon)
  - Searching for mlncRNA's
  - mlncRNA examples in plants
- CNEs (Abiy)
- Heterochromatin (Patrick)



# What is Dark Matter?

- Comings (1972) was the first person to use the term "junk DNA"

- How much is there?

- 99% of human genome doesn't encode proteins at all

- Difficult to estimate % functional DNA

- **50 bp-windows:** parameter dependent

- **Partitioning of bases:** parameter and method dependent

- **Critics:** assumptions in above 2 models cause significant underestimation

- **Indel model:** model of indels instead of nucleotide substitutions

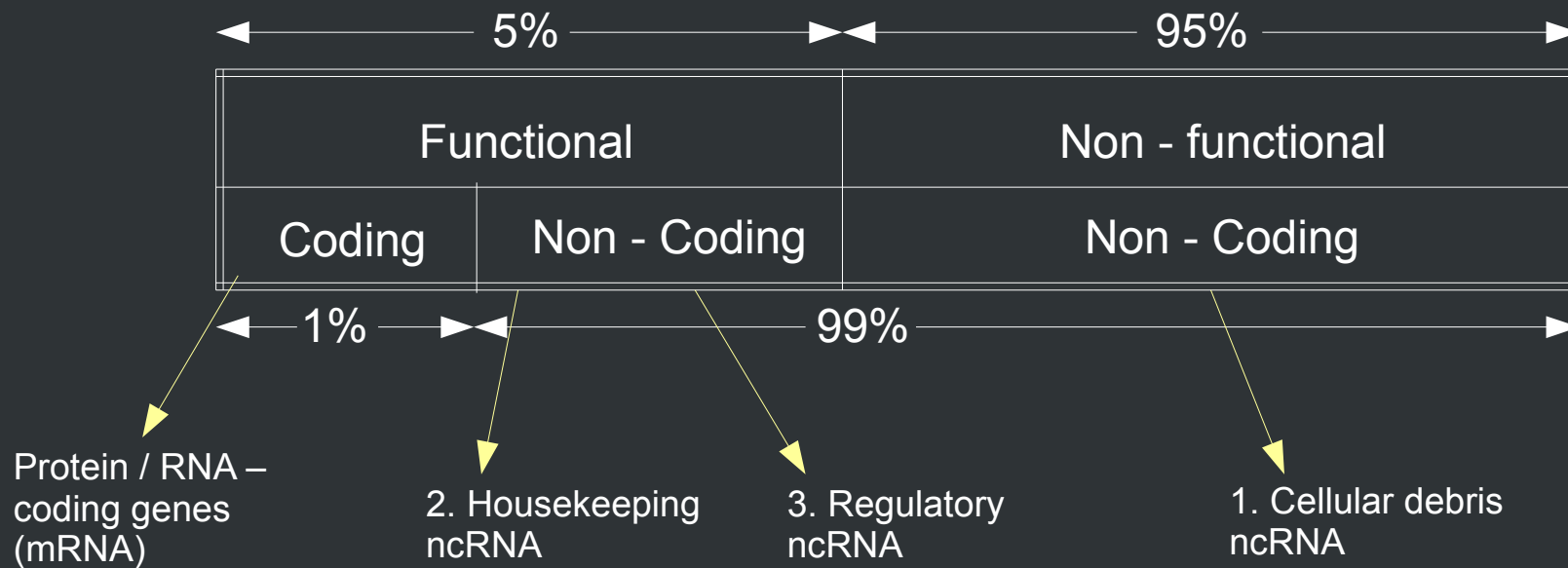
- 5% functional / 2% - 8% functional / other predictions  $\gg$  5% functionality

- $\pm$  3% of human euchromatin is functional



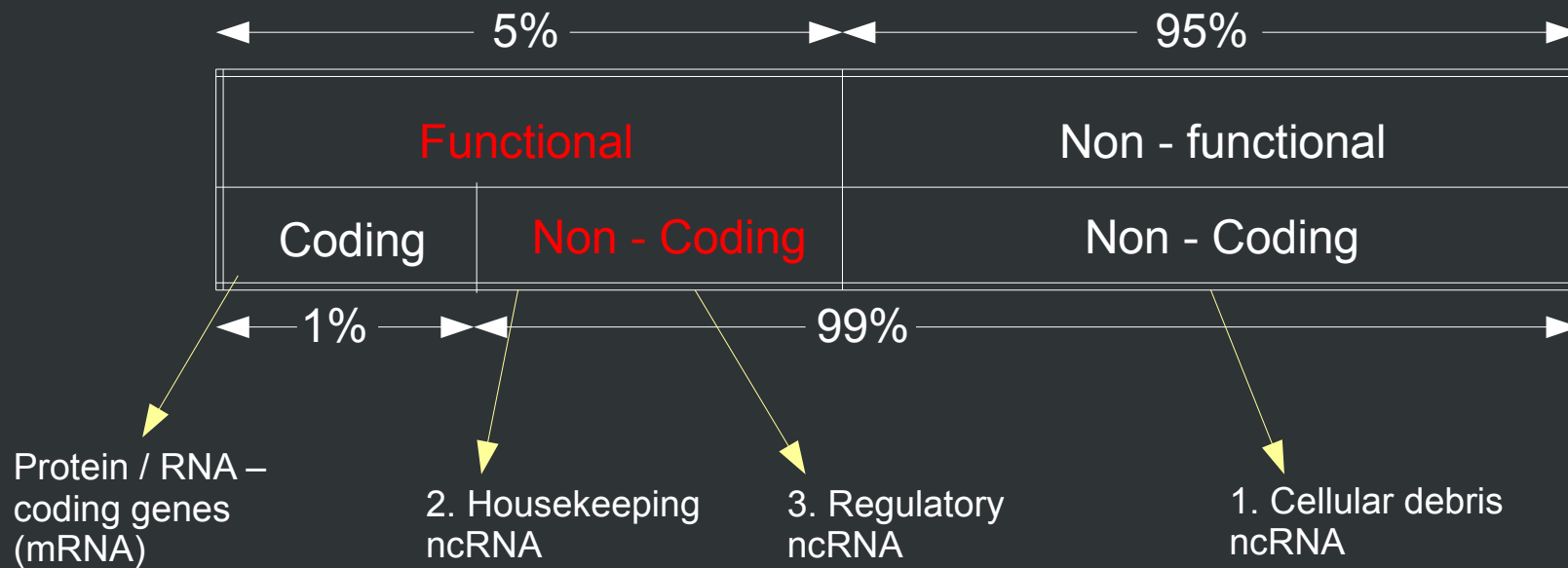
# What is Dark Matter?

- Putting it into perspective



# What is Dark Matter?

- Putting it into perspective



# What is Dark Matter?

## 1. Cell debris ncRNA

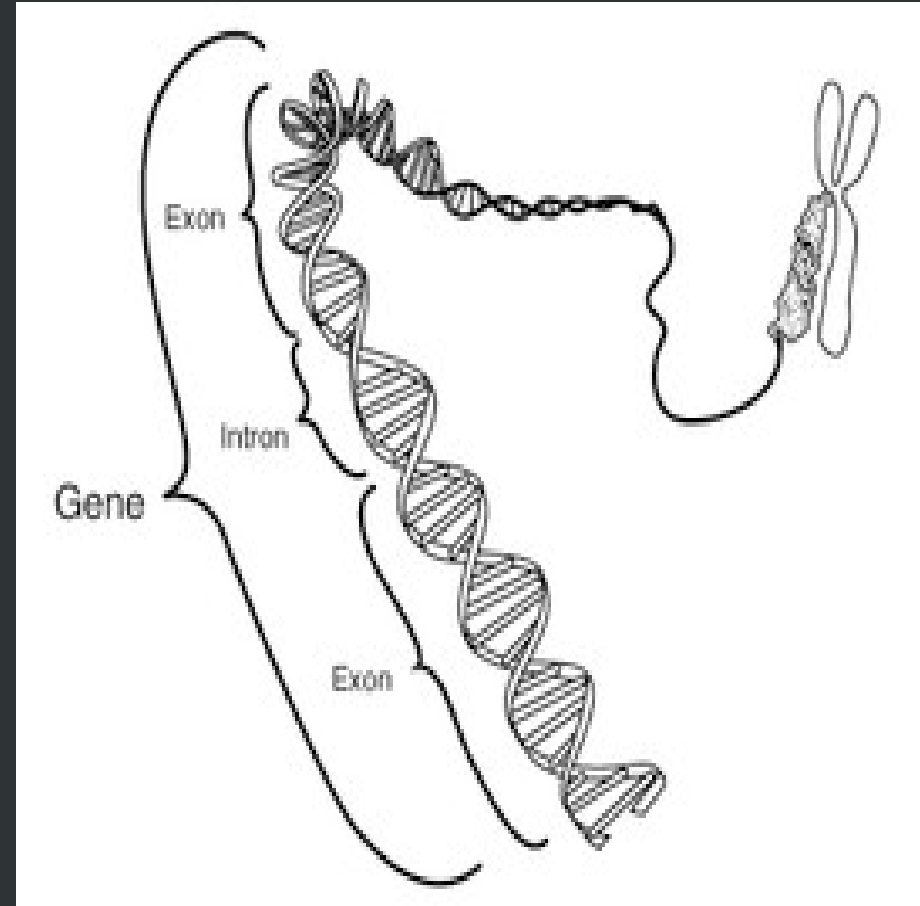
Found mostly in introns and pseudogenes:

- Temporary classification
- No apparent function

## 2. Housekeeping ncRNA

Examples:

- tRNA
- rRNA
- small nuclear/nucleolar RNA (sRNA/snRNA)
- signal recognition particle (7SL/SRP)RNA



# What is Dark Matter?

## 3. Regulatory ncRNA

Two categories:

- Small (short)      i.e. 20 – 40 nucleotides
- Large (long)      i.e. > 40 nucleotides

(i) Small:

- examples:      heterochromatic siRNA, long siRNA, trans-acting siRNA, natural antisense siRNA, natural antisense miRNA, CNEs
- function:      regulates gene expression through DNA methylation, mRNA decay or translational inhibition



# What is Dark Matter?

## 3. Regulatory ncRNA

Two categories:

- Small (short)      i.e. 20 – 40 nucleotides
- Large (long)      i.e. > 40 nucleotides

(i) Small:

- examples: **heterochromatic siRNA**, long siRNA, trans-acting siRNA, natural antisense siRNA, natural antisense miRNA, **CNEs**
- function: regulates gene expression through DNA methylation, mRNA decay or translational inhibition



# What is Dark Matter?

## 3. Regulatory ncRNA

(i) Large:

-- examples:    miRNA and several others (CNEs incl.)

-- function:    (animals) control imprinting and dosage  
                  compensation of the X chromosome, modulation of  
                  transcription and translation

(plants) studied less, but appears to regulate the  
phosphate-starvation response, gender specific  
expression, and nodulation



# What is Dark Matter?

## 3. Regulatory ncRNA

(i) Large:

-- examples: **mlncRNA** and several others (**CNEs** incl.)

-- function: (animals) control imprinting and dosage compensation of the X chromosome, modulation of transcription and translation

(plants) studied less, but appears to regulate the phosphate-starvation response, gender specific expression, and nodulation



# Plant Dark Matter

- Searching for mlncRNAs:

cDNA – selected from size-selected RNA or ESTs

GENOME – overlaps coding genes?

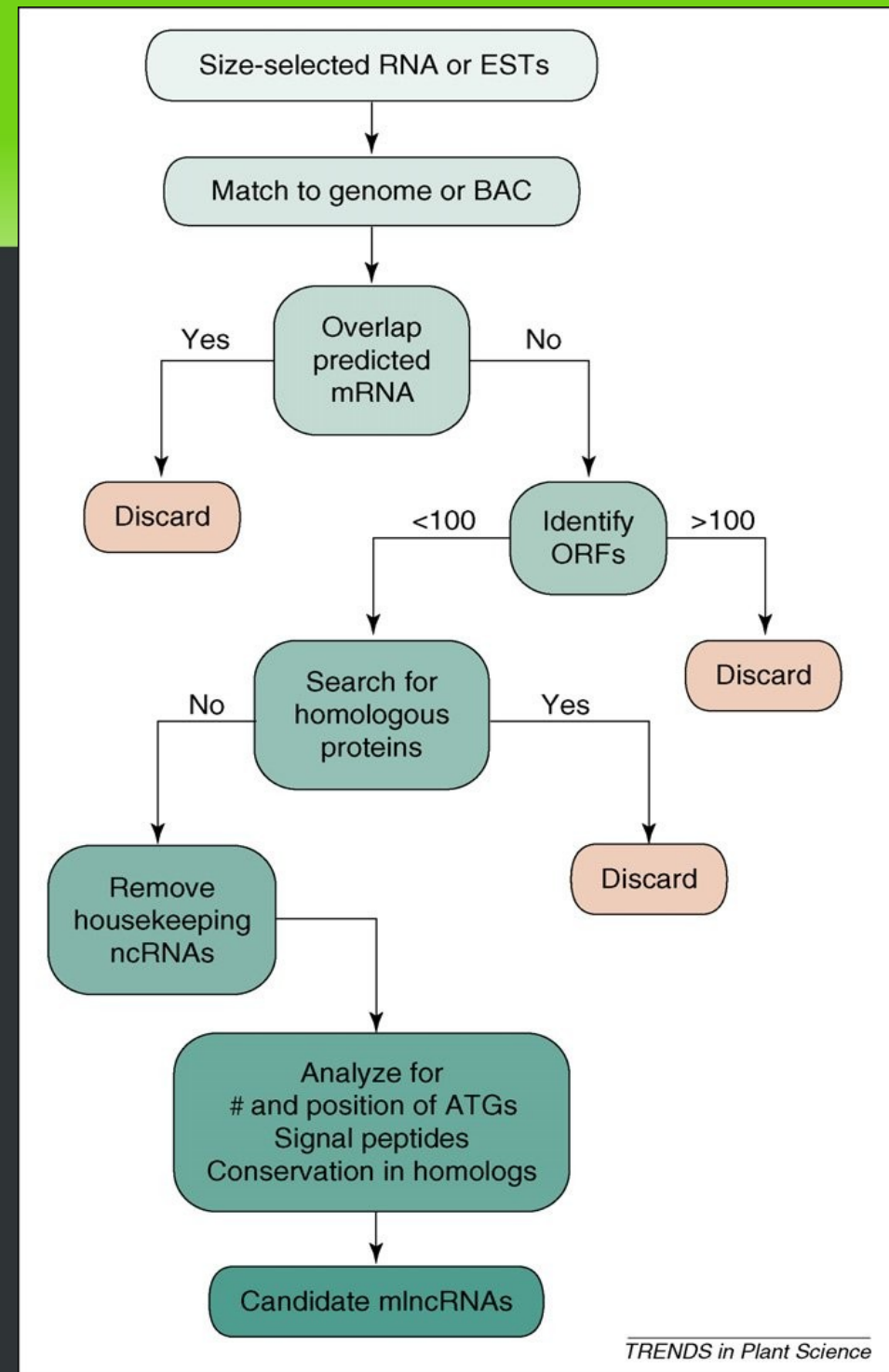
ORFs – >100 amino acid discarded

HOMOLOGY – homology to known proteins, housekeeping genes or if there are any repeats discard

## WHAT IS LEFT ???

pepRNA and mlncRNAs

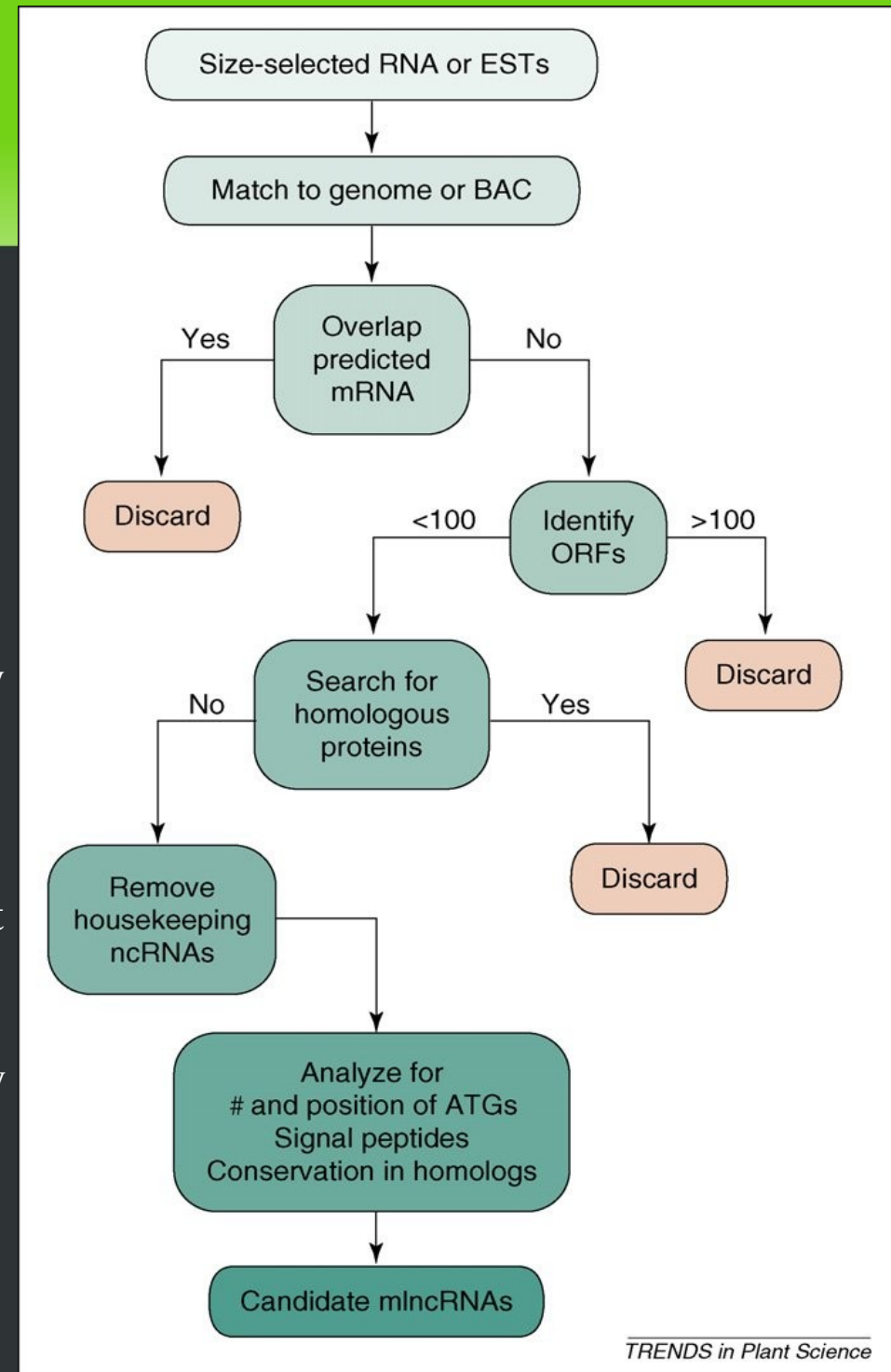
- How can we distinguish between them?



# Plant Dark Matter

## • Searching for miRNAs:

- DIFFERENCES – pepRNA has very few ATGs upstream of the predicted ORF, miRNAs have numerous
- pepRNA ORFs are more likely to contain signal peptides
  - pepRNAs have mutation predominantly in the third position of their codons, whilst miRNA can be at any position provided crucial sequence motifs and secondary structures are maintained



# Plant Dark Matter

- miRNA examples in plants

- Natural Antisense Transcripts (NATs)

- miRNAs & response to phosphate deprivation

e.g. TPSI1/Mt4 gene family

- Gender-associated miRNAs in plants

e.g. CsM10, CCLS96.1 & Zm401

- Nodulation

e.g. Enod40 gene family

- Hammerhead ribozymes

*Arabidopsis thaliana* & *Medicago truncatula*



# Plant Dark Matter

## Natural Antisense Transcripts (NATs):

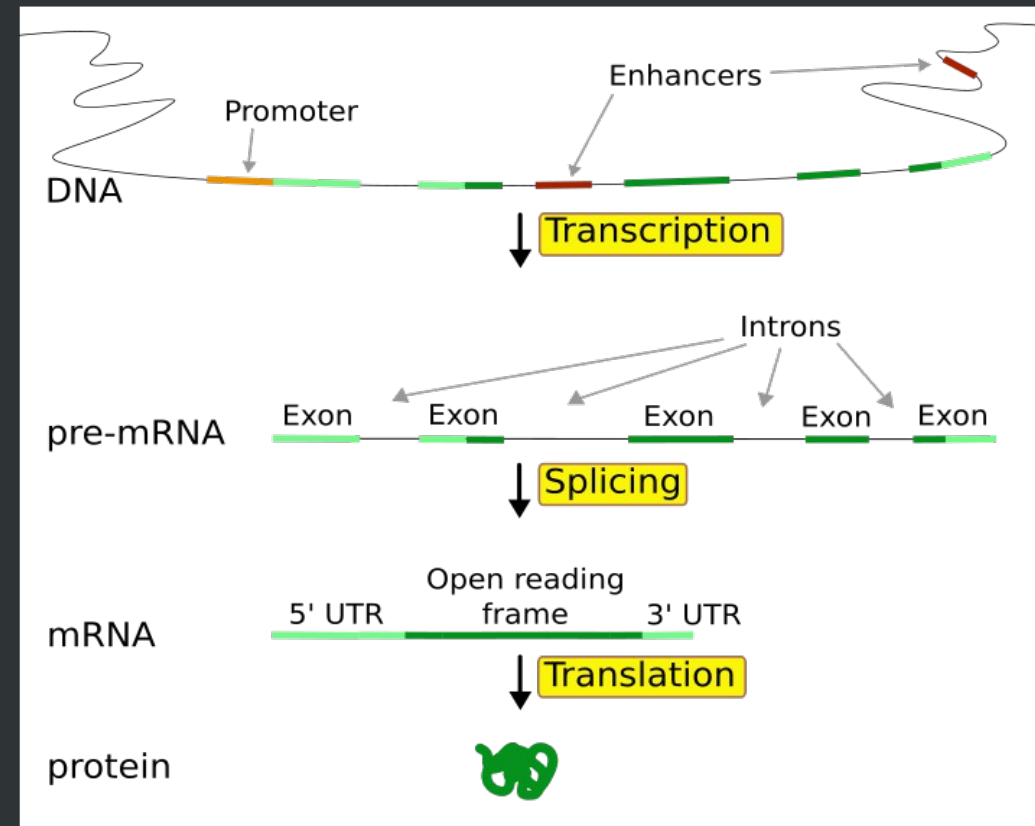
- often discarded in mlncRNA search due to homology with mRNA
- significant portion of the transcriptome
  - (13% have coding potential  $\leq 100$  amino acids)
  - => pepRNA or mlncRNA
- function: regulation of splicing & polyadenylation



# Plant Dark Matter

## miRNAs & response to phosphate deprivation: (TPSI1/Mt4) gene family

- TPSI/Mt4 released under phosphate starvation conditions
- miRNA399 upregulated
- miRNA399 guides a complex of proteins to relevant target (PHO2), initiates cleavage resulting in reduction of phosphate signalling
- 24 nucleotide sequence conservation
- function of TPSI1/Mt4: prevents miRNA399 cleavage to target – why?



**FINE TUNING  
OF  
RESPONSE**



# Plant Dark Matter

## Gender-associated mlncRNAs in plants:

- 11% flowering plants produce unisex flowers, mlncRNA expressed in two ways
  - (i) preferentially in one gender
  - (ii) spacially and temporally dependent
- CsM10 (belongs to CGR mlncRNA family)
  - function: CGR2 instability in tobacco ensures that RNA accumulation is highly responsive to environmental / developmental conditions  
=> RNA transcription moderation



# Plant Dark Matter

## Gender-associated mInRNAs in plants:

### - CCLS96.1

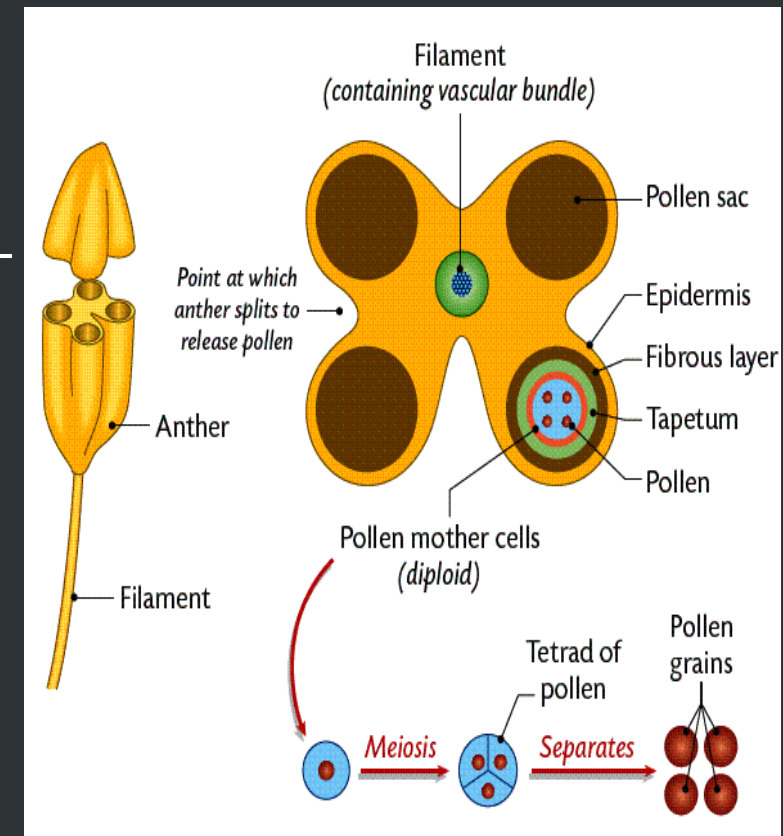
-- % is 3-fold higher in male flowers, but 4% - 8% higher in female leaves

-- function: as yet undetermined

### - Zm401

-- overexpression & suppression reduces viability & amount of pollen produced

-- function: regulation of the timing of tapetum degradation during pollen development



# Plant Dark Matter

## Nodulation (Enod40 gene family):

- function: positively/negatively influences ethylene levels in *Medicago* & tobacco respectively which influences nodule formation

suggestion made: function dependent on location (i.e. tissue type, species & experimental conditions)



Root nodule

## Hammerhead ribozymes:

- function: unknown as yet



Hammerhead ribozyme



# Intermission Break...

- Please feel free to take a brief moment to stretch your legs... Abiy will be ready to start again in a few moments on his topic entitled

“CNEs”



# Conserved Non-Coding Elements(CNEs)

- The human genome contains thousands of non-coding sequences that are often more conserved between vertebrate species than protein-coding exons
- Method
  - Comparing human genome against distantly related vertebrates.
  - Indicate
    - Appearance of CNEs which have been frozen throughout evolution



# CNEs...

- Example.
  - Human, mouse and Rat share 256 elements, 100% identical, over > 200bp
  - Human and Japanese pufferfish share 1373 CNEs, 199bp and 84% identity
- Location
  - Clustering in genomic regions that contain gene coding for transcription factors and signaling genes involved in the regulation of development ('trans-dev genes')
- Function
  - Act as cis-regulatory sequences for these trans-dev genes.



# Relation of Vertebrate and invertebrate CNEs.

- Using sequence similarity hard to trace CNEs in invertebrates.
  - Unknown evolutionary origin of vertebrates CNEs
- Recently, 20,301 CNEs identified with >50bp between *D.melanogaster* and *D.pseudoobscura*.
- Location
  - preferentially near genes encoding transcription factors and developmental regulatory genes



# Challenges Faced

- distinguishing functionally conserved elements from background sequence conservation by comparing these two genomes alone
- unclear how widespread highly conserved non-coding elements are among different animal genomes and whether similar genes are associated with the most conserved non-coding elements in both invertebrate and vertebrate genomes.

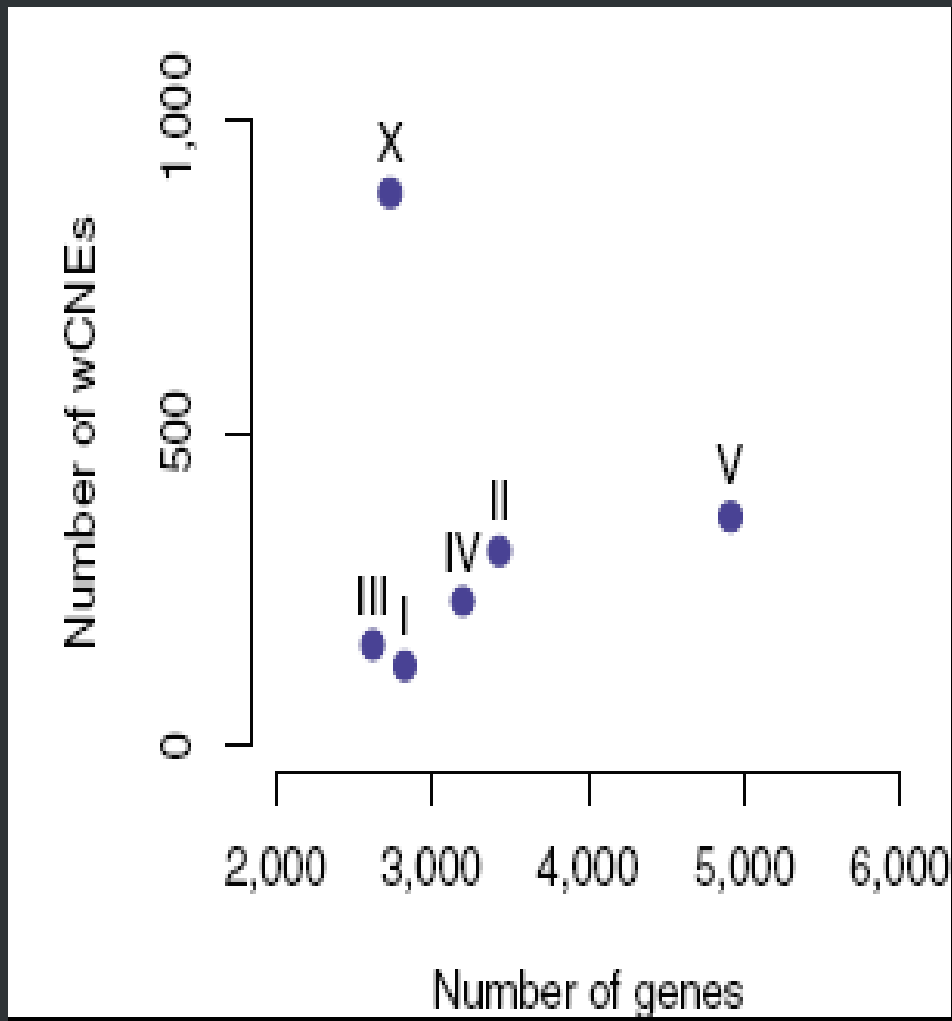


# CNEs for nematode Genomes

- Using Mega Blast to search for sequences between genomes of *C.elegans* and *C.Briggsae*
  - Large block of identity
  - No evidence of transcription
- Found 3,601 with 30 identity sequences
- 69% also in *C.remanei*, 1460, intergenic (no transcription), 624 in introns.
- Parameters related with vertebrate CNEs
  - Location
  - A+T content



# Location (along chromosomes)



- residing in the gene-rich center of autosomes, and multiple of them clustering around single gene.
- Example:-
  - 42% are found on single C.elegans sex chromosome.

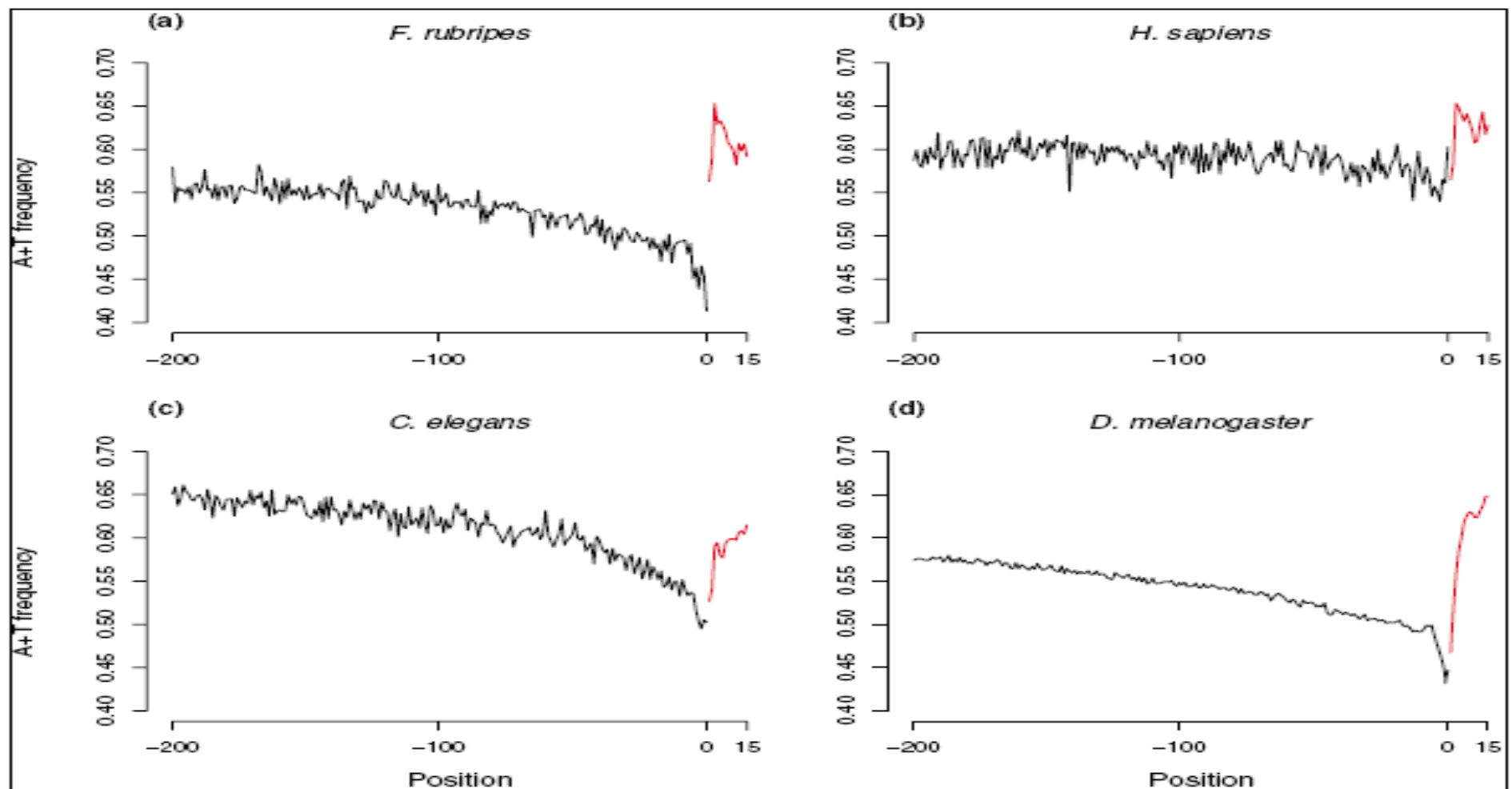


# Contd.....

- Implying wCNEs act as cis-regulatory architectures for this chromosome since it is almost devoid of essential genes.
- Nucleotide (A+T) composition at boundaries
  - A gradual G+C enrichment followed by a sharp A+T rich peak at the CNE bridges mark transition of base composition from DNA to CNE DNA.



# Contd....



**Figure 2**  
CNEs share a striking nucleotide signature from *C. elegans* to vertebrates. The plot shows the percentage of A+T nucleotides for 200 bp of sequence flanking CNEs (black) and 15 bp of CNE (red) at the CNE border defined by sequence conservation (the sequence on one end of each CNE is reverse complemented) for (a) *F. rubripes*, (b) *H. sapiens*, (c) *C. elegans* and (d) *D. melanogaster*. In all four species there is a decrease of A+T content in the 200 bp of sequence flanking the CNEs followed by a sharp A+T increase at the CNE border.

# Association with trans-dev genes

- Associating each wCNEs to the protein-coding gene with nearest transcription start site.
  - Identified CNE- associated genes are related to transcription factor activity and development.



# Genes associated with wCNEs

- enriched for cell-signaling terms .
  - Less striking in humans (except for major signaling genes involved in development, eg; Sonic hedge-hog gene)
- enriched in the neighbourhoods of genes encoding DNA-binding transcription factor domain.



# Orthologs

- Of 397 human CNE-associated genes 190 have identifiable orthologs in *C.elegans*.
- 60/190 associated with wCNEs of *elegans*.
- 40/60 orthologs in *Drosophila* associated with its CNEs.
  - 40/156 human CNE-associated genes have orthologs in both *C.elegans* and *D.Melanogaster*.
    - Represent core set of developmental regulatory genes that are associated with CNEs across three different phyla.



# Function hypothesis

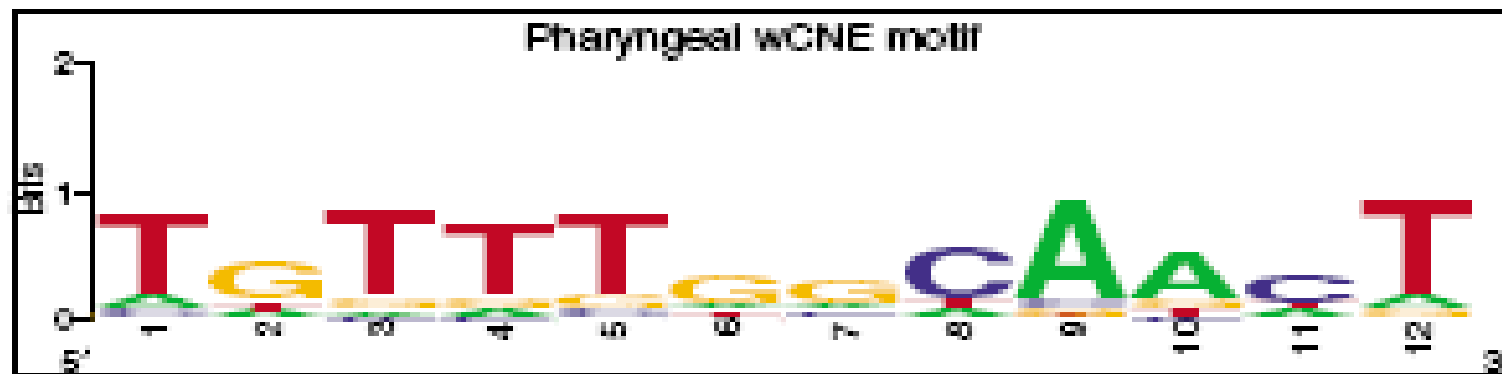
- CNEs they encode arrays of transcription factor binding sites.
  - =>CNEs associated with genes known to be expressed in a particular tissue type should be enriched for DNA-binding sites for transcription factors regulating the coexpression of these genes in that tissue.
- Test
  - 54 wCNE-associated genes in C.elegans Pharynx
    - Associated with 120 wCNEs
      - 40 intronic, 80 intergenic



# Method

- Weeder motif discovery algorithm
  - Search for overrepresented motifs
  - Post process to identify similar('redundent') motifs among high scoring motifs.
- Result of algorithm
  - A motif significantly enriched in this sequences
    - Motif very similar to the consensus binding site of the pharyngeal transcription factor PHA-4.(major specifier of pharyngeal cell identity in C.elegans.(pha-4/FoxA is critical to establish the onset of transcription of Caenorhabditis elegans foregut (pharynx)) genes
    - Suggests occurrence of this motif in wCNEs represent genuine PHA-4 binding sites.





Occurrences of a sequence motif overrepresented in wCNEs associated with pharyngeal genes

wCNE coordinates	Strand	Matching sequence	Position	Score	wCNE distance from TSS	Gene name
IV:3776258..3776298	+	TATTTAGCATCT	9	85.59	9,435	<i>vab-2</i>
IV:8369551..8369581	-	TTTTTTGCAACT	3	91.65	347	D2096.6
V:10673732..10673841	-	TGTTTGTCCACT	15	87.26	1,202	<i>ceh-22*</i>
V:13217316..13217419	+	TGTTTGGCAACT	23	100	3,588	F57B1.6
X:2215856..2215898	-	TGTTTGA AATT	12	85.67	230	<i>peb-1</i>
X:6621897..6621968	-	TTTATGGCAACT	47	88.99	826	C25B8.4
X:7457940..7457992	+	TGTTTGACAATT	5	91.56	2,212	<i>sox-2</i>

We used the Weeder motif discovery program to search for overrepresented motifs in all wCNEs spatially associated with genes predicted to be expressed in the pharynx based on microarray data [31]. From this dataset, Weeder identified a motif very similar to the consensus binding site for PHA-4, the master specifier of pharyngeal cell identity (TRTTKRY, where R = A/G, K = T/G, and Y = T/C) [33, 34]. This table shows the coordinates (WormBase version WS140) of the wCNEs that contain matches to the overrepresented motif, the coordinates of the matches within the wCNEs, the Weeder scores of the matches to the motif, the distances (in bp) between the wCNEs and the transcription start site (TSS) of the associated genes and the names of the associated genes. The predicted site in the element 1.2 kb upstream of *ceh-22* (marked with an asterisk) lies within a 30 bp pharyngeal muscle enhancer bound by PHA-4 [35].

# Conclusion

- CNEs are associated with genes involved in
  - transcription regulation
  - development
  - Cell -signaling ( to a certain degree) in both vertebrates and invertebrates.
- Function
  - WCNEs represent enhancer sequences that function by encoding transcription factor binding sites.



# Intermission Break #2...

- Please feel free to take a brief moment to stretch your legs... Mr Phepa will be ready to start again in a few moments on his topic entitled  
“Heterochromatins”



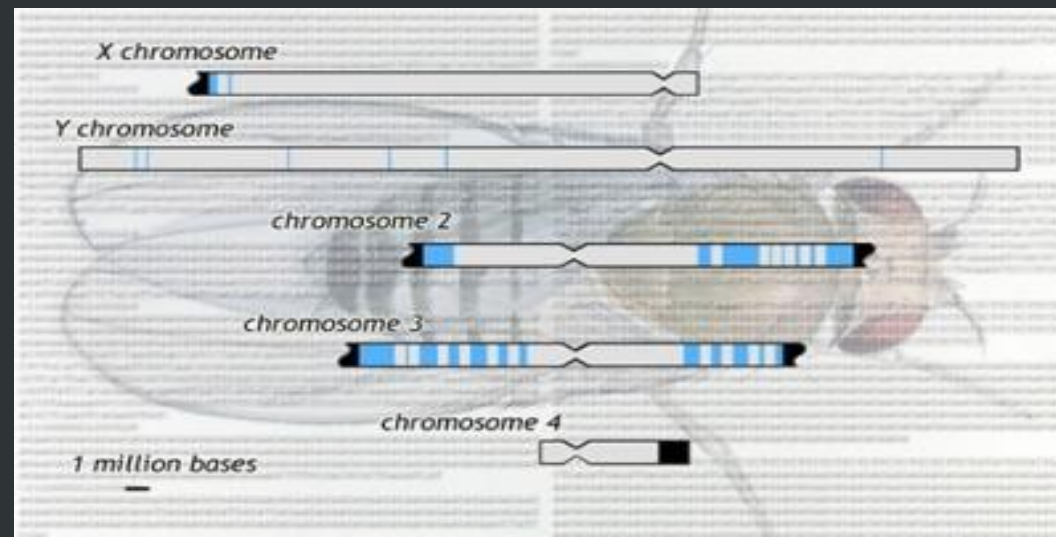
# Dark Matter Heterochromatin

- One of the component of the highly repetitive gene-poor DNA(Dark matter).
- The interest to study the heterochromatin was developed by a project called DHGP.
- The results found confirmed that heterochromatin is far from mere junk.



# Thoughts of scientists about Heterochromatin

- Heterochromatin had little or no function. This was so because it lacked the protein-coding genes that occur so richly in the chromosomes' more accessible areas called euchromatin.
- But for the past years it has become apparent that heterochromatin is critical for a number of essential functions.



# Advances on sequencing Drosophila Heterochromatin

- The advances brought the following to light:
  - Extended understanding of the heterochromatin's organization and constitution.
  - Led to new insights into how it helps cells and organisms survive.



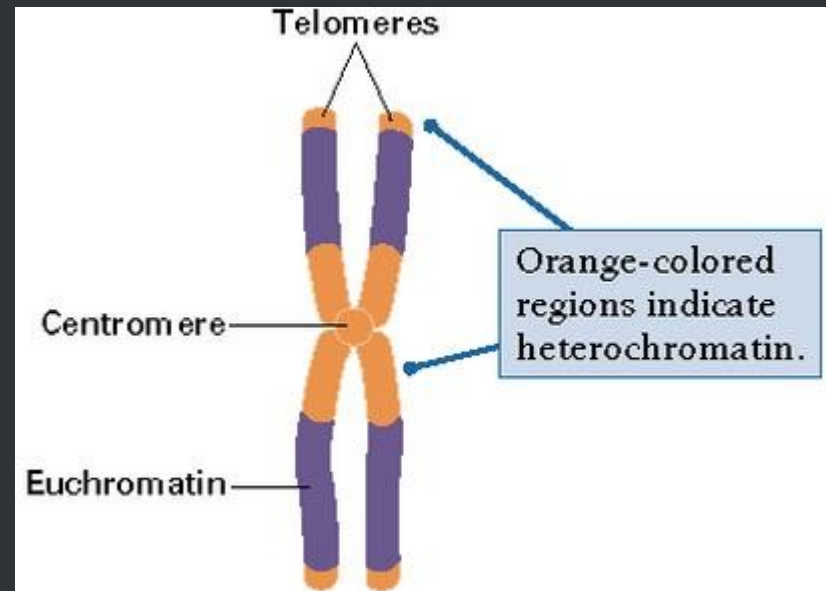
# Latest Results

- Published by DHGP in June 15,2007.
- Reviewed that in the heterochromatin regions ,there are 200 protein-coding genes.
- The heterochromatin also includes other features of biological importance.
  - Non-protein-coding RNAs and other functional elements like small RNAs were also discovered.
  - These small RNAs neutralise transposable elements- DNA similar to viruses that hop around the genome and are capable of disrupting gene function.



# Regions of Heterochromatin

- Found in the centromeres and telomeres
- Centromeres play a crucial role in controlling chromosomeduplicatin during cell division.
- Telomeres help prevent the accumulation of genomic damage.



# Method used to study Drosophila genome

- Whole-genome shotgun sequencing. In the method, the whole fly is crushed and produce libraries of DNA fragments of different lengths eg 1000 bases.
- Then the matching of the sequences takes place. This proves a huge job to do because of the repetitive nature of the heterochromatin.



# Hallmark of heterochromatin

- Repeating sequences, and are in different distinct kinds.
  - Simple , short repeats are called “satellite DNA”- are more abundant near the centromeres.
- In the “seas” of the satellite DNA, there are islands of moderate- length repeats made up of transposable elements.



# Concluding Remarks

- The studies led to the understanding of the dead region of the genome.
- Also reviewed that proper understanding of the whole genome is completed with the knowledge of heterochromatin.



# Concluding Remarks

- In a nutshell all the remarks are encapsulated in the following metaphor by Smith:
  - “We don't know what holds the galaxies together, and the same is true of the genome. We are pretty good at understanding how individual genes work, but we do not understand, for example how the large-scale structures of genomes affects cellular processes. We hear too much about the “post-genomic era”-it is underappreciated that we don't understand the genome yet”



# References

- [http://en.wikipedia.org/wiki/RNA\\_interference](http://en.wikipedia.org/wiki/RNA_interference)
- The functional repertoires of metazoan genomes
  - Chris P. Ponting
- Diamonds in the rough: mRNA-like non-coding RNAs
  - Linda A Raymonds, James P. Kastenmayer, Alexander G. Huttenhofer and Pamela J. Green



# References

- Parallel evolution of conserved non-coding elements that target a common set of developmental regulatory genes from worms to humans.
  - Tanya Vavouri, Klaudia Walter, Walter R Gilks, Ben Lehner and Greg Elgar
- Statistical information characterization of conserved non-coding elements in vertebrates.
  - I.Abnozova, K.Walter, R. te Boekhorst, G.Elgar, W.R.Gilks



# References

- Sequence finishing and mapping of *Drosophila Melanogaster* Heterochromatin.
  - Roger A.Hoskins et al
- The release 5.1 annotation of *Drosophila melanogaster* heterochromatin.
  - Christopher D.Smith et al

