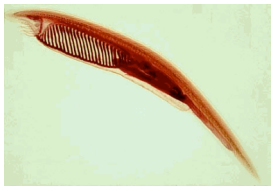
Evolutionary genomics of the recently duplicated amphioxus Hairy genes

Senda Jiménez-Delgado[1], Miguel Crespo[2], Jon Permanyer[1], Jordi Garcia-Fernàndez[1#] and Miguel Manzanares[2#]

Sean McAllister, Emily Loter, Aaron Gibbs

What is an amphioxus?

- Amphioxus = both ends pointed, in Greek
- *Branciostoma lanceolatum* is the most common and well known of the approximately 25-30 species
- Typical beach bums…they can usually be found in shallow, tropical and temperate oceans and spend most of their time buried in the sand
- Part of the Cephalochordate branch of the animal kingdom
- Grow to be 5-8cm long at adulthood
- There are separate sexes, and eggs are fertilized externally, which then develop into free-swimming larvae
- They don't have any hard parts, and as a result, their fossil record is very sparse



Amphioxus

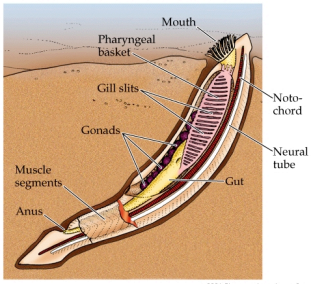
The closest invertebrate relative to the vertebrates

Possesses a vertebrate like body plan:

- ***Notochord*** – a muscularized nod that supports the nerve chord
- ***Hollow dorsal nerve chord***
- ***Segmented muscle blocks*** – these are V-shaped structures called myomeres
- ***Perforated pharyngeal region*** – also called "gill slits," which strain food particles out of the water
- ***Post anal tail***

Devoid of the most complex features of vertebrates

- ***Elaborate brain*** – the amphioxus brain is very small and poorly developed, as are most of its sensory organs
- ***Paired fins***
- ***No true vertebrae***



Evolution

- The amphioxus is not extinct, and in fact is enjoying a life living on the temperate sands of Jamaica (among other sandy shores)
- It is believed to be the closest living invertebrate relative to vertebrates
- Its genome has been independently evolving, as has the genome of vertebrates, since their divergence

Amphioxus
Branchiostoma lanceolatum

A fish-like thing appeared among the annelids one day.
It hadn't any parapods nor setae to display.
It hadn't any eyes nor jaws, nor ventral nervous cord.
But it had a lot of gill slits and it had a notochord.

(chorus):
It's a long way from Amphioxus. It's a long way to us.
It's a long way from Amphioxus to the meanest human cuss.
Well, it's goodbye to fins and gill slits, and it's welcome lungs and hair!
It's a long, long way from Amphioxus, but we all came from there.

It wasn't much to look at and it scarce knew how to swim.
And Vertis was very sure it hadn't come from him.
The mollusks wouldn't own it and the arthropods got sore.
So the poor thing had to burrow in the sand along the shore.

So burrowed in the sand before a crab could nip his tail.
And he said "Gill slits and myotomes are all to no avail.
I've grown some metapleural folds and spent an oral hood.
But all these fine new characters don't do me any good.

(chorus)

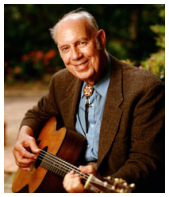
It sulked awhile down in the sand without a bit of pep.
Then he stiffened up his notochord and said, "I'll beat 'em yet!
Let 'em laugh and show their ignorance. I don't mind their jeers.
Just wait until they see me in a hundred million years.

My notochord shall turn into a chain of vertebrae
And on fins my metapleural folds will agitate the sea.
My tiny dorsal nervous cord will be a mighty brain
And the vertebrates shall dominate the animal domain.

(chorus)

Amphioxus song

Folk Musician and Marine Biologist Sam Hinton wrote and performed this song, which is sung to the tune, "It's a long way from Tipperary."

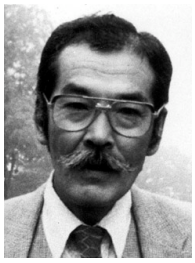


Garcia-Fernández J. Amphioxus: a peaceful anchovy filled to illuminate Chordate Evolution (I). Int J Biol Sci 2006; 2:30-31. http://www.biolsci.org/v02p0030.htm

The 2R Hypothesis and DDC Model

Formulation of the 2R hypothesis

- Susumu Ohno (1928-2000)
- Wrote *Evolution by Gene Duplication (1970)*
- Hypothesized that there were two or more full genome duplications in the early evolution of vertebrates
- He bases his hypothesis solely on genome size in different chordates and evidence of tetraploidization in fish lineages

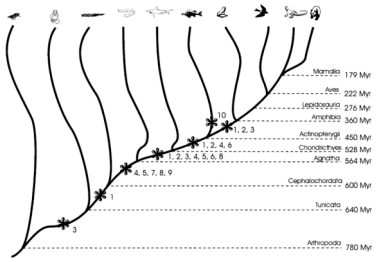


Formulation of the 2R hypothesis

- In the last decade, the "two rounds of genome duplication in vertebrate evolution hypothesis (2R hypothesis)" resurfaced and gained popularity among biologists
- A series of articles propose different numbers and timing for the whole genome duplication
- Hughes, et al. devised a number of experiments to test whether or not the 2R hypothesis was valid.

The Popular Vote

- The most popular version of the 2R hypothesis:
  - one duplication event at the root of the vertebrate lineage
  - another around Agnatha and Gnathostomata.
- Some extreme proposals include the first genome duplication at the root of Chordata lineage and the last one at the amphibian lineage.



A schematic representation phylogeny of main invertebrate groups.

Timescales are based on molecular clock (Kumar and Hedges 1998). Some of the proposed vertebrate whole genome duplications are marked by asterisks. The connected asterisks denote the range of the proposed genome duplication time. The references are as follows: 1, Ohno 1970; 2, Ohno 1970; 3, Lundin 1993; 4, Holland et al. 1994; 5, Sidow 1996; 6, Kasahara et al. 1996; 7, Spring 1997; 8, Ohno 1998; 9, Meyer and Schartl 1999; 10, Amores et al. 1998 and Mayer and Schartl 1999.

Arguments for the 2R Hypothesis

- Gene numbers in different animals lineages vary greatly
- Many invertebrates have ~15,000 genes
- With the initially accepted human gene number ~80,000, the fourfold ratio between mammalian and invertebrate gene numbers was in good agreement with the 2R hypothesis.

Houston, we have a problem…

- The current mammalian gene number estimations based on both ESTs and draft sequence of the human genome reveal that our genome hosts much fewer protein coding genes than anticipated
- The 35,000 genes in the human genome means that, on average, for every invertebrate protein gene there are only two mammalian orthologs.

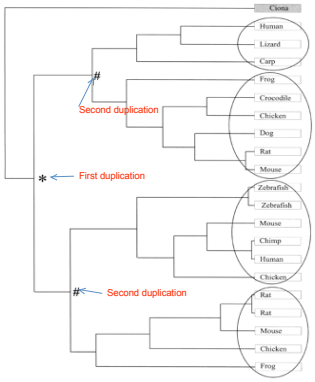
Arguments for the 2R Hypothesis, continued

- One can argue that many redundant genes have been lost during vertebrate evolution.
- If this were true, we should be able detect gene loss by simply plotting phylogenetic trees for different vertebrate gene families.

A hypothetical phylogenetic tree of vertebrate gene family according to the 2R hypothesis.

•Two rounds of the whole genome duplication would result in four vertebrate gene clusters of (AB) (CD) tree topology. This topology should be easily detectable even in incomplete data or after a gene loss in specific lineages.

•This figure represents a hypothetical phylogenetic tree for a vertebrate gene family, assuming some gene loss or incomplete data and additional gene duplication in specific lineages.



2R, or not 2R

- To support the 2R hypothesis, clustered paralogous genes not only have to exhibit (AB) (CD) tree topology, but additionally they ought to diverge in concordance with the *Hox* family and by the proposed 2R hypothesis genome duplication time.
- The investigators (Hughes, et al) applied four different methods:
  - construction of phylogenetic trees
  - estimation of the divergence time of paralogous genes
  - consistency of the phylogenetic trees
  - the parsimony test
- The results of their analysis did not support the 2R hypothesis. Thirty-five gene families provided evidence regarding the hypothesis, and in 29 cases the results were inconsistent with the hypothesis.
- For more information on the specifics of their tests, please see the following link: Are we Polyploids? A Brief history of one hypothesis

Maybe not for now…

- Formulated three decades ago, the hypothesis of whole genome duplications in the early stages of vertebrate evolution has had as many adherents as opponents.
- It seems that the current data do not support the 2R hypothesis, and the existence of more paralogs in vertebrates than in invertebrates can be explained by waves of tandem duplications of single genes or larger chromosomal fragments.

Evolution is tricky…

- Reconstruction of the history of living organisms is a very difficult task.
- We are not able to reconstruct it with certainty because of its complexity.
- Many evolutionary events become obscure with time
  - hence, inferences about the early evolution of vertebrate genomes remain in a scientific "gray zone"
- It is probable that we never will be able to say *how this happened* but only *how it could have happened*.

The DDC Model

DDC = duplication, degeneration, complementation

- The origin of organismal complexity is generally thought to be tightly coupled to the evolution of new gene functions arising subsequent to gene duplication.
- Under the classical model for the evolution of duplicate genes, one member of the duplicated pair usually degenerates within a few million years by accumulating deleterious mutations, while the other duplicate retains the original function.
- This model further predicts that on rare occasions, one duplicate may acquire a new adaptive function, resulting in the preservation of both members of the pair, one with the new function and the other retaining the old.
- However, empirical data suggest that a much greater proportion of gene duplicates are preserved than predicted by the classical model.

DDC Model

- The duplication-degeneration-complementation (DDC) model predicts:
  - degenerative mutations in regulatory elements can increase rather than reduce the probability of duplicate gene preservation
  - the usual mechanism of duplicate gene preservation is the partitioning of ancestral functions rather than the evolution of new functions

What this has to do with Amphioxus…

- In the vertebrate lineage, the evolution from the last ancestor most probably involved two rounds of genome duplication
- This increase in the number of genes was likely instrumental, by means of sub-functionalization of gene duplicates
- Specifically, the acquisition of vertebrate novel features, such as the migratory neural crest and derivatives, skeletal system, cartilaginous tissue and a complex brain

What these scientists are using the DDC for in this paper:

- Duplicated copies of a single gene suffer from differential loss of *cis*-regulatory regions.
- Now, a complex or pleiotropic function that was performed by a single gene prior to duplication, is now subdivided into discrete components.
- These copies are now all very necessary and essential, as they keep individual and unique *cis*-regulatory regions.

The Amphioxus Genome

- Has also been evolving since the separation from its last common ancestor with vertebrates
- We expect to find cephalocordate-specific duplicates for some gene families
- This has been reported, and the extreme case is to be found in the *Hairy* family of helix-loop-helix transcription factors

What is a *Hairy* gene?

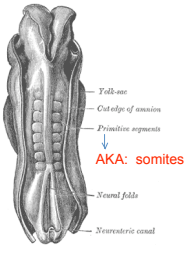
Hairy genes play an important role in the developmental processes of many organisms

- Formation of somites
- In the developing vertebrate embryo, somites are masses of mesoderm distributed along the two sides of the neural tube and that will eventually become dermis, skeletal muscle, and vertebrae.

Somite examples

Chick embryo somites

Human embryo somites



Helix-Loop-Helix Transcription Factors

- Amphioxus – Eight members (*Hairy A*–*Hairy H*)
- Mouse – One member (*Hes1*)
- Chicken – Two members
- *Xenopus* (frog) – Two members
- Zebrafish – Two members

Amphioxus *Hairy* Genes

- The 8 amphioxus *Hairy* genes conserve their gene structure without excessive accumulation of mutations, expansions or losses, and have no stop codons within the coding sequence
- This indicates that they have not degenerated to pseudogenes, and they should be expressed at some point throughout the amphioxus life cycle
- Four of them have known expression during development (*HairyA-D*)

HairyA – HairyD

- Expressions are mostly non-overlapping, or overlapping in a subtle manner during particular developmental stages
- From the expression data, it has been proposed that amphioxus *Hairy* genes, after gene duplication from a single ancestor, underwent a process of divergence in the *cis*-regulatory regions that matches the DDC model.

HairyA – HairyD

- Although *HairyA-HairyD* have specific expression and minor overlapping, the overall regions or tissues where they are expressed are the same.
- Expression is found in the neural tube, presomitic mesoderm, somites, endoderm or notochord
- Therefore, the researchers expected to find conserved elements in the regulatory sequences of these four genes

How these particular *Hairy* genes might function in amphioxus:

These conserved elements could represent enhancers with spatial information for these somitic regions, refining gene expression.

Notch Signaling Pathway

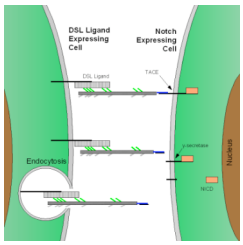
- In zebrafish, chicken and mice, various factors involved in the formation of somites are under the control of the Notch signaling pathway.
- The RBP-Jk binding sites are the primary transcriptional mediators of Notch signaling
- One of these targets in mice is *Hes1*, a *Hairy* gene which contains RBP-Jk sites in its 5' regulatory region.

Details of the Notch Pathway

1. Maturation of the Notch receptor involves cleavage at the prospective extracellular side during intracellular trafficking in the Golgi complex.
2. This results in a bipartite protein, composed of a large extracellular domain linked to the smaller transmembrane and intracellular domain.
3. Binding of the ligand promotes two proteolytic processing events.
4. As a result of proteolysis, the intracellular domain is liberated and can enter the nucleus to engage other DNA-binding proteins and regulate gene expression.

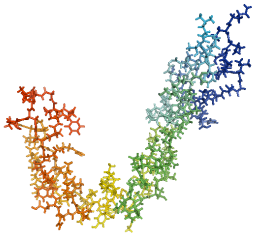
Notch Signaling Cascade

1. Once the Notch extracellular domain interacts with a ligand, the enzyme *TACE* cleaves the Notch protein just outside the membrane.
2. This releases the extracellular portion of Notch, which continues to interact with the ligand.
3. The ligand plus the Notch extracellular domain is then endocytosed by the ligand-expressing cell.
4. There may be signaling effects in the ligand-expressing cell after endocytosis; this part of Notch signaling is a topic of active research.
5. After this first cleavage, an enzyme called γ-secretase cleaves the remaining part of the Notch protein just inside the inner leaflet of the cell membrane of the Notch-expressing cell.
6. This releases the intracellular domain of the Notch protein, which then moves to the nucleus where it can regulate gene expression by activating the transcription factor CSL.
7. Other proteins also participate in the intracellular portion of the Notch signaling cascade.



Notch-1 Mechanism

The Notch protein sits like a trigger spanning the cell membrane, with part of it inside and part outside. Ligand proteins binding to the extracellular domain induce proteolytic cleavage and release of the intracellular domain, which enters the cell nucleus to alter gene expression.



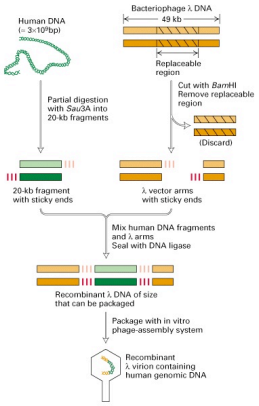
Goals of the Paper

- Gain insights into the process of amplification of the *Hairy* family in Amphioxus
- Identify *cis*-regulatory modules responsible for the maintenance of the duplicates
- Decipher whether amphioxus *Hairy* genes may be under the control of the Notch pathway

How they did it…

- First, they cloned the full coding region of *HairyA-F* plus 3kb of 5' regulatory regions into lambda phage
- Then, an *in silico* analysis of potential *cis*-regulatory control elements was performed

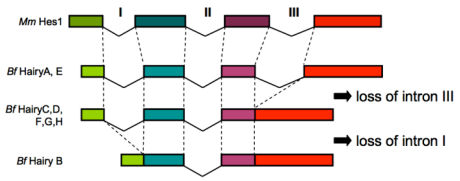
Engineering a lambda phage using a human DNA insert



Genomic Organization

- Compared genomic sequence (from lambda phage clone libraries) to predicted peptides to determine exon/intron structure
- Compared these results to *Hairy* genes in mice (*Hes1*), chicken (*Hairy1*), and the urochordate *Ciona intestinalis*

Genomic Organization



- Chick *Hairy1* has all three introns
- *Ciona intestinalis* (urochordate)
  - *hairya* and *hairyb* have all three introns
  - *hairyc* lacks intron three
- *Bf* = Amphioxus; *Mm* = *Mus musculus*

Intron Loss after *Hairy* Gene Duplication Events

- Given that Amphioxus *HairyA* and *E*, mouse *Hes1*, chick *Hairy1*, and *Ciona intestinalis hairya* and *b* all contain three introns, it is most likely that this exon/intron structure represents that of the common ancestor to chordates
- In the Amphioxus line:
  - Intron III was lost in *B,C,D,F,G,* and *H*
  - Intron I was lost in *B*

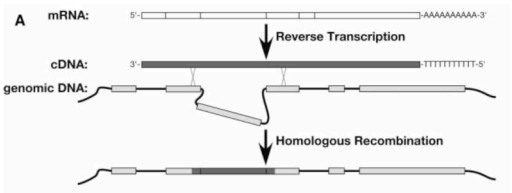
Is intron loss common?

- It is not as infrequent as one might think—researchers* studying intron loss in *Caenorhabditis* found that it had a very high rate of selective intron loss (nearly 400-fold higher than loss rates in mammals)

*Cho S, Jin SW, Cohen A, Ellis RE. A phylogeny of caenorhabditis reveals frequent loss of introns during nematode evolution. Genome Res 2004; 14: 1207-1220.

Mechanisms of Intron Loss

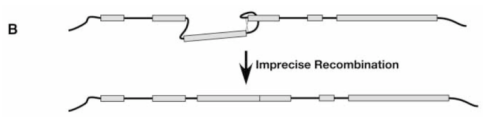
- mRNA mediated recombination



Cho et al. Genome Res 2004; 14: 1207-1220.

Mechanisms of Intron Loss

- mRNA mediated recombination
- non-homologous recombination between short repeats at the 5' and 3' ends of introns

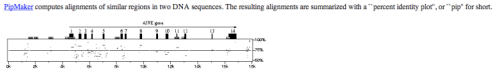


Cho et al. Genome Res 2004; 14: 1207-1220.

Comparative Genomic Methods

- Jiménez-Delgado et al. (2006) used the following tools for identifying conserved regions in/around the *Hairy* gene:

Pipmaker



http://pipmaker.bx.psu.edu/pipmaker/

Comparative Genomic Methods

- Jiménez-Delgado et al. (2006) used the following tools for identifying conserved regions in/around the *Hairy* gene:

Mulan

http://mulan.dcode.org/

Comparative Genomic Methods

- Jiménez-Delgado et al. (2006) used the following tools for identifying conserved regions in/around the *Hairy* gene:

Vista

http://genome.lbl.gov/vista/index.shtml

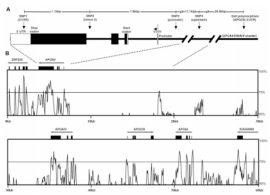
We have seen Vista Plots Before…

An Apolipoprotein Influencing Triglycerides in Humans and Mice Revealed by Comparative Sequencing

Len A. Pennacchio,[1] Michael Olivier,[2*] Jaroslav A. Hubacek,[3] Jonathan C. Cohen,[3] David R. Cox,[2] Jean-Charles Fruchart,[4] Ronald M. Krauss,[1] Edward M. Rubin[1†]

Comparison of genomic DNA sequences from human and mouse revealed a new apolipoprotein (*APO*) gene (*APOAV*) located proximal to the well-characterized *APOA1/C3/A4/A5* gene cluster on human 11q23. Mice expressing a human *APOAV* transgene showed a decrease in plasma triglyceride concentrations to one-third of those in control mice; conversely, knockout mice lacking *Apoav* had four times as much plasma triglycerides as controls. In humans, single nucleotide polymorphisms (SNPs) across the *APOAV* locus were found to be significantly associated with plasma triglyceride levels in two independent studies. These findings indicate that APOAV is an important determinant of plasma triglyceride levels, a major risk factor for coronary artery disease.

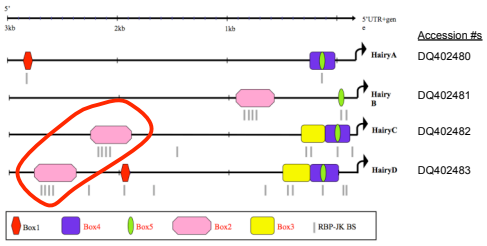
www.sciencemag.org SCIENCE VOL 294 5 OCTOBER 2001



They say…

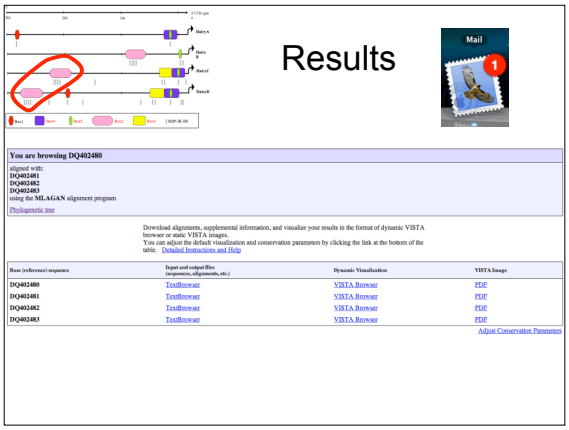
- VISTA is a comprehensive suite of programs and databases for comparative analysis of genomic sequences.
- mVISTA is a set of programs for comparing DNA sequences from two or more species up to megabases long and visualize these alignments with annotation information.

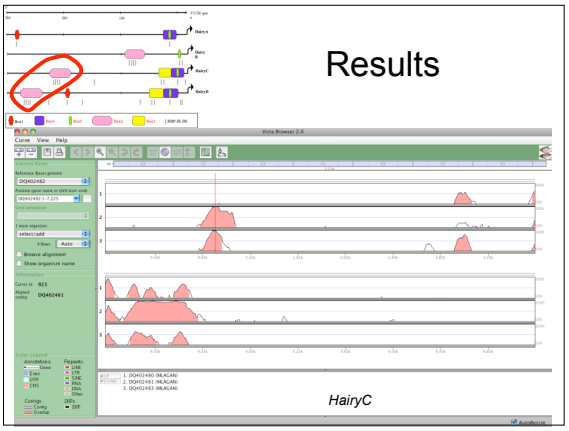
A Practical Example

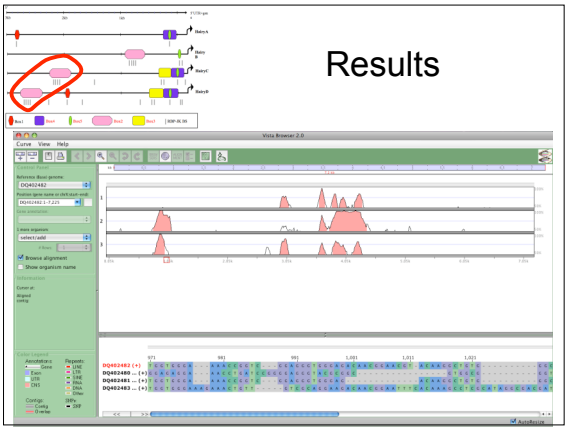


GenBank sequence includes the *Hairy* gene plus 3kb of 5' upstream sequence

BEGIN HERE





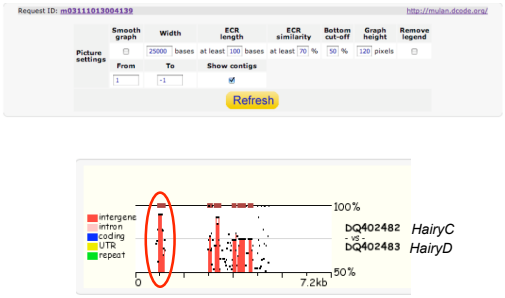


Can Support Hypotheses with Mulan (and Pipmaker)

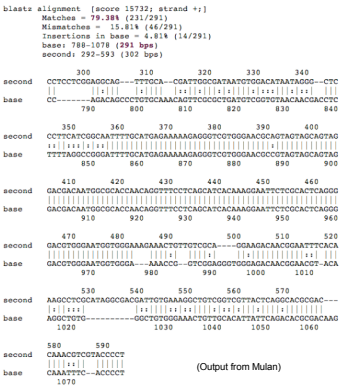
- Mulan does essentially the same thing as Vista:

local sequence alignment

Mulan Results



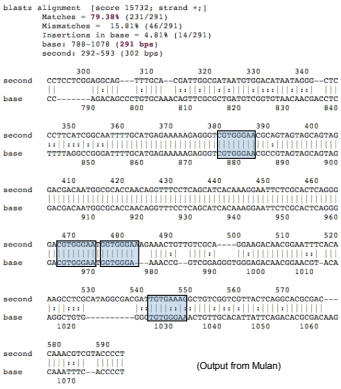
Both Vista and Mulan came up with the following sequence for this region of high conservation



This is the same conserved "box" that the researchers found

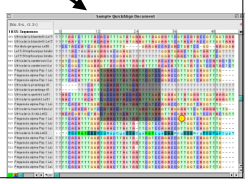
Can use Matinspector to find transcription factor binding sites

Blue boxes highlight RBP-Jk binding sites



Remember:

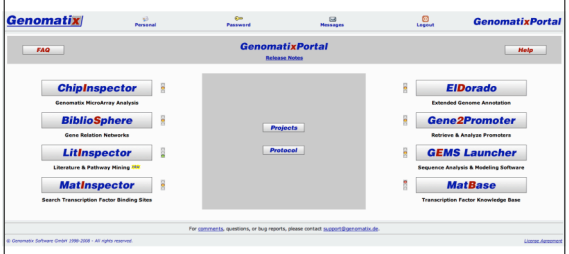
- After comparing their sequences in Vista, Pipmaker, and Mulan, the researchers ran a more complicated sequence alignment algorithm using Clustal X
- This was done before running sequences through Matinspector



Location of Transcription Factor Binding Sites Using Matinspector

- Find upstream regulators of *Hairy* genes
  - Evolutionarily conserved between *Hairy A-D*
    - Like RBP-JKappa
- Matinspector
  - A "software tool that utilizes a large library of matrix descriptions for transcription binding sites to locate matches in DNA sequences."
  - Matinspector

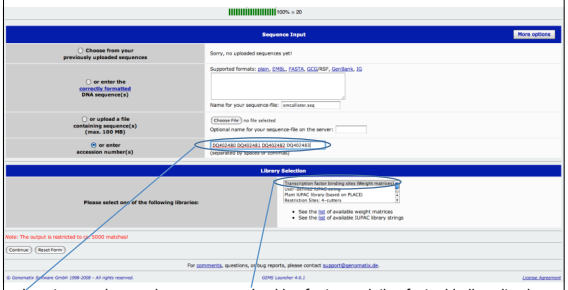
MatInspector



Opening page
→Click MatInspector

Many other tools for similar searches

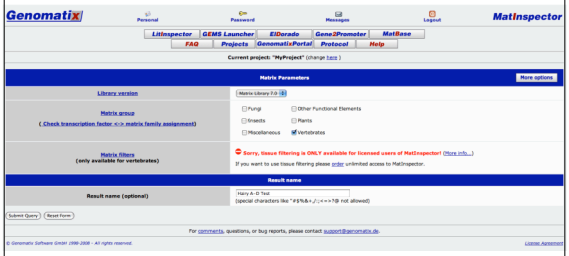
MatInspector



Insert accession numbers
-also can insert DNA sequences

Looking for transcription factor binding sites in entire gene

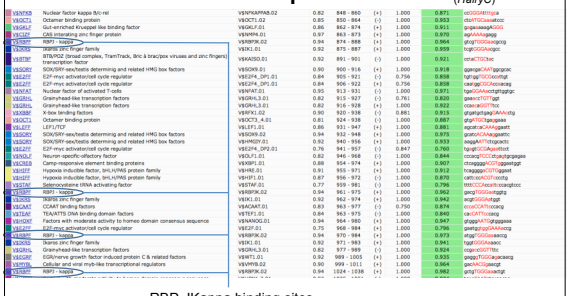
MatInspector



View in many groups

MatInspector

(*HairyC*)



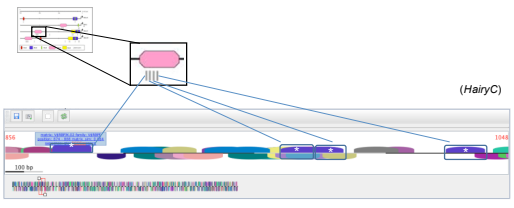
RBP-JKappa binding sites

1 page of a very long list of transcription factor binding sites

MatInspector

-The RBP-JKappa Binding sites are found in the gene

-Reflect direct correlation of RBP-JKappa found in Box 2 in *HairyB-D* genes

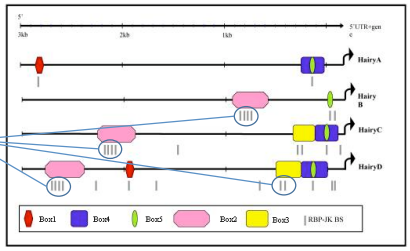


(*HairyC*)

RBP-JKappa Binding Sites

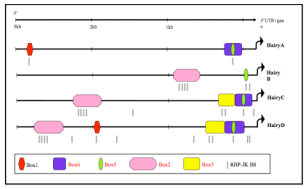
- RBP-JKappa binding sites are primary transcriptional mediators of Notch signaling pathway
  - Presence of RBP-JKappa BS in non-coding 5' region of Hairy genes
    - Illustrates functional role of regulation of Hairy genes during development
    - Shows Hairy genes are downstream targets of Notch pathway

5’ Non-coding Region of Amphioxus Hairy Genes



RBP-JK BS present in conserved regions

Most likely *Hairy B,C,& D* have related function during development



D

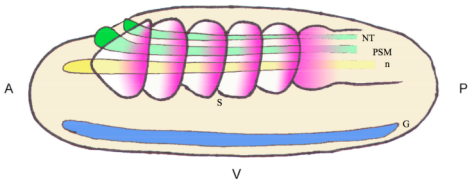
Box1: unknown for now

Box2: somites & presomitic mesoderm

Box3: notochord

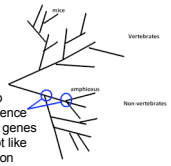
Box4: gut endoderm

Box5 (ie: core of Box4): neural tube



Summary

- Proving DDC model in amphioxus
  - Difference in Boxes → different expression of genes in different tissues
    - Example: Expression of HairyC & D in notochord
- Observed differential gene expression during development
  - In particular Hairy genes
- Set out to connect genomes of amphioxus to vertebrates



Due to divergence points genes are not like common ancestor