Title
Logic, genomics, and evolution of the peripheral nervous system transcriptional network of Drosophila

Permalink
https://escholarship.org/uc/item/76w4b7hh

Author
Rebeiz, Mark John

Publication Date
2006

Peer reviewed|Thesis/dissertation
UNIVERSITY OF CALIFORNIA, SAN DIEGO

Logic, genomics, and evolution of the peripheral nervous system

transcriptional network of Drosophila

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biology

by

Mark John Rebeiz

Committee in charge:

Professor James W. Posakony, Chair
Professor Randall S. Johnson
Professor Christopher R. Kintner
Professor William McGinnis
Professor Pavel A. Pevzner

2006
The dissertation of Mark John Rebeiz is approved, and it is acceptable in quality and form for publication on microfilm:

Chair

University of California

2006
For my wife, Tarcia K. Rebeiz
# TABLE OF CONTENTS

Signature Page .................................................................................................................. iii

Dedication ............................................................................................................................ iv

Table of Contents ............................................................................................................... v

List of Figures and Tables .................................................................................................... vi

Preface ................................................................................................................................... ix

Acknowledgements ............................................................................................................. x

Vita, Publications, and Fields of Study ............................................................................. xiv

Abstract ............................................................................................................................ xvii

Introduction ........................................................................................................................ 1

Chapter 1: GenePalette: a universal tool for genome sequence navigation
  and visualization ............................................................................................................... 19

Chapter 2: SCORE: A computational approach to the identification of *cis*
  regulatory modules and target genes in whole genome
  sequence data.................................................................................................................... 28

Chapter 3: Notch regulates *numb*: integration of conditional and
  autonomous cell fate specification mechanisms............................................................. 37

Chapter 4: A code for neural precursors............................................................................. 77

Chapter 5: An ancient transcriptional regulatory linkage................................................. 117
# LIST OF FIGURES AND TABLES

**Introduction**

Figure I.1: Bristle organs of *Drosophila melanogaster* ........................................ 9

Figure I.2: The Notch pathway acts at multiple levels of bristle development. ................................................................. 12

**Chapter One**

Figure 1: Accessing sequence data in GenBank using GenePalette ........... 22

Figure 2: The GenePalette interface provides three perspectives of a Genome sequence segment in a single window.......................... 23

Figure 3: Integration of the Graphical View with other Interface components ................................................................. 24

Figure 4: Interconnectivity of GenePalette interface components .............. 25

**Chapter Two**

Figure 1: Global inventory of Su(H) binding site clusters in the *Drosophila* genome ........................................................................ 31

Figure 2: Purity of Su(H) binding site cluster bins in the fly genome........ 32

Table 1: Su(H) binding site clusters identified by SCORE .......................... 33

Figure 3: Sequences including a cluster of four Su(H) binding sites Between *Her* and *Him* are required for normal expression *Him* ........................................................................................................ 34

Figure 4: Su(H) binding sites are highly significantly clustered in the fly genome............................................................................. 34
Chapter Three

Figure 3.1: Validation of putative Su(H) binding sites in numb found by whole genome in silico analysis ........................................ 64

Figure 3.2: numb is specifically expressed in the bristle lineage ............... 65

Figure 3.3: Restriction of Su(H) responsive bristle lineage enhancer to a 683 bp fragment active in pIIa and pIIIb cells ......................... 65

Figure 3.4: Notch signaling is required for activity of the CD2 reporter fragment ................................................................. 67

Figure 3.5: Alignment of the CD2 regulatory region across nine Drosophila species ........................................................................ 68

Figure 3.6: Expression of mutant CD2 GFP reporters .................................. 69

Figure 3.7: Model for numb expression in the pIIa cell ................................... 70

Table 3.S1: Table 3.S1 Primers used in the study ........................................ 71

Figure 3.S1: Expression of CD1 and CD2 fragments in embryos .................. 72

Chapter Four

Figure 4.1: A hypothetical dual input model for an SOP enhancer module... 100

Figure 4.2: A cluster of proneural and bHLH-repressor sites in the intron of phyllopod .................................................................. 101

Figure 4.3: Binding of proneurals and bHLH-repressors to E-Boxes within the phyllopod intron .................................................. 102

Figure 4.4: Expression of phyllopod intron reporters in 3rd instar wing imaginal discs ................................................................. 103

Figure 4.5: A cluster of proneural and bHLH-repressor sites in the nervy locus ................................................................. 104
Figure 4.6: An SOP enhancer of nvy restricted by bHLH-R binding sites ..... 105

Figure 4.S1: Survey of multiple wildtype phyl GFP lines............................... 106

Figure 4.S2: Survey of multiple repressor mutant phyl GFP lines................. 107

Figure 4.S3: Survey of multiple proneural mutant phyl GFP lines.............. 108

Figure 4.S4: Expression of phyl reporter in the absence of
achaete and scute .................................................................................................................... 109

Table 4.S1: Primers used in this study to create EMSA probes
and reporter constructs........................................................................................................ 110

Table 4.S2: Clusters of proneural + bHLH-repressor sites identified by
SCORE in the same significance bin as phyl.................................................... 111

Chapter Five

Figure 1: Conservation of Hes-type bHLH repressor (bHLH-R) sites
among proneural genes........................................................................................................ 121

Figure 2: A phylogeny of proneural genes............................................................ 122

Figure 3: Evolutionary conservation of bHLH-R DNA-binding specificity .... 124

Figure S1: Replicate of phylogenetic tree shown in Fig. 2, with
branch lengths and support values included......................................................... 128

Figure S1(cont’d): Replicate of phylogenetic tree shown in Fig. 2, with
branch lengths and support values included ........................................ 129
PREFACE

Seven years ago, as a research technician in a cattle genomics lab, I decided to apply to graduate programs. We were doing a pre-genomic sequence project of comparatively mapping the cattle genome to the then unsequenced mouse and human genomes. It was during this time that I learned to program. It was an exciting period, and I was braced for the things we could do once the vertebrate genomes were sequenced. While contemplating what I wanted to do, I was drawn to model organisms like *Drosophila melanogaster*, *Caenorhabditis elegans*, and even *Saccharomyces cerevisiae*. It was exhilarating to see the promise of their whole genome sequence, and I was awestruck by the power these systems offered experimentally. I read with great excitement papers with titles such as “Life with 6000 genes”, and the *Drosophila* genome sequence issue of *Science* magazine. These experiences strongly influenced my thoughts on the future of genomic biology as I was writing personal statements for graduate school applications.

Now, seven years later, I am proud to present my thesis, which embodies my own participation in this post-genomic era movement.
Acknowledgements

This work would not have been possible without the advice and supervision of Jim Posakony. In addition to his technical guidance in the field of developmental biology, I have benefited tremendously from his rigorous and critical approach to scientific research, the creative license that he has afforded me, and his constant vision of where a field is going (a vision that more often than not, points North).

I have been fortunate to share the Posakony lab with many talented and intelligent individuals. I thank them for making the lab an inviting, and scientifically stimulating environment. I am indebted to Nick Reeves, Scott Barolo and Par Towb for their constant help as I started working in the lab. I am also thankful for the input and dialog with current lab members: Brian Castro, Steve Miller, and Joe Fontana. I also thank Tammie Stone for her expert assistance with gel shifts.

I must acknowledge the input and close interactions with many members of the McGinnis lab. I am particularly indebted to Dave Kosman, who has selflessly devoted hours to helping me with confocal microscopy. I have enjoyed many discussions with Joseph Pearson, Derek Lemmons, and Matthew Ronshaugen concerning genome tools, and bioinformatic approaches. As a member of 4th floor, Bonner Hall, I must also thank Jackie
Vignes for her expert help with administrative matters, and her ever-supportive nature.

Thanks to my mom and dad, who have been so supportive during my scientific career. The weekly Sunday phone calls with my parents were almost as motivating as lab meeting for finding new data – my Dad has asked me how the research is going on each of the 312 Sundays of my graduate career. To my in-laws, I must say thank you for the constant support, and especially all of the free baby-sitting.

To my children, Kezia and Zander, I must say thank you for being a major source of motivation and smiles. To my wife, Tarcia, I say thanks for following me across the country, the constant support, encouragement, and for all of the times I have been late without a call.

Chapter One, in full, is a reprint of the following publication. Rebeiz M, Posakony JW. GenePalette: a universal software tool for genome sequence visualization and analysis. *Dev Biol.* 2004 Jul 15;271(2):431-8. I was the primary researcher and author and James W. Posakony directed and supervised the research that forms the basis for this chapter.

Chapter Two, in full, is a reprint of the following publication. Rebeiz M, Reeves NL, Posakony JW. SCORE: a computational approach to the

I am the primary researcher and author and James W. Posakony directed and supervised the research that forms the basis for this chapter. Characterization of the *Him* gene through GFP reporter and *in situ* hybridization experiments were performed by Nick L. Reeves.

Chapter Three, in full, is a manuscript in preparation for publication:
Notch regulates numb: integration of conditional and autonomous cell fate mechanisms. Rebeiz M, Stone T, and Posakony JW. I am the primary researcher and author and James W. Posakony directed and supervised the research that forms the basis for this chapter. Tammie Stone performed the electrophoretic mobility shift assay that is presented in Figure 3.1.

Chapter Four, in full is a manuscript in preparation for publication:
Lateral inhibitory logic revealed: *phyllodopod* is a target of both proneural activators and bHLH-repressors. Rebeiz M, Stone T, Posakony JW. I am the primary researcher and author and James W. Posakony directed and supervised the research that forms the basis for this chapter. Tammie Stone performed the electrophoretic mobility shift assays in Fig. 4.3.
Chapter Five, in full, is a reprint of the following publication. Rebeiz M, Stone T, Posakony JW. An ancient transcriptional regulatory linkage. *Dev Biol.* 2005 May 15;281(2):299-308. I am the primary researcher and author and James W. Posakony directed and supervised the research that forms the basis for this chapter. Tammie Stone performed the electrophoretic mobility shift assays.
VITA

1973  Born, Champaign, Illinois

1995  B.S., University of Illinois at Urbana-Champaign

1996-1998 Research Assistant
       Animal Sciences
       University of Illinois at Urbana-Champaign

1999  M.S. University of Illinois at Urbana-Champaign

1998 – 2000 Research Specialist in the Lab of Harris Lewin,
          University of Illinois at Urbana-Champaign

2000 - 2006 Teaching and Research Assistant
          Division of Biological Sciences
          University of California, San Diego

2006  Ph.D., University of California, San Diego

PUBLICATIONS

Rebeiz M, Stone T, Posakony JW. An ancient transcriptional regulatory

Everts-van der Wind A, Kata SR, Band MR, Rebeiz M, Larkin DM, Everts RE,
Green CA, Liu L, Natarajan S, Goldammer T, Lee JH, McKay S, Womack JE,
Lewin HA. A 1463 gene cattle-human comparative map with anchor points
defined by human genome sequence coordinates. *Genome Res.* 2004
Jul;14(7):1424-37.

Rebeiz M, Posakony JW. GenePalette: a universal software tool for genome


**FIELDS OF STUDY**

**Major field: Biology**

Studies in Developmental Genomics  
Professor James W. Posakony  
University of California, San Diego

Studies in Comparative Genomics  
Professor Harris A. Lewin  
University of Illinois at Urbana-Champaign
Studies in Reproductive Biology
Professor David J. Miller
University of Illinois at Urbana-Champaign
ABSTRACT OF DISSERTATION

Logic, genomics, and evolution of the peripheral nervous system
transcriptional network of Drosophila

by

Mark John Rebeiz
Doctor of Philosophy in Biology
University of California, San Diego, 2006

Professor James W. Posakony, Chair

With the advent of whole genome sequences, biologists are confronted with the problem of understanding the non protein-coding portion of the genome. This remains a difficult task due to the many enigmatic features of cis-controlling sequences. In this thesis, I will present the genomic tools, methodologies, and observations I have made in the course of studying the peripheral nervous system of Drosophila melanogaster, a model system well suited for the study of transcriptional regulatory networks.

In the first chapter, I will describe a software tool that I created to facilitate the use of genome sequences in wet-lab experiments. This tool,
GenePalette, is particularly well suited for inspecting genomic regions that have been implicated by whole-genome analysis (searches for transcription factor binding sites, core promoter sequences, microRNA target sites, etc), and it was used extensively in the analyses presented in each subsequent chapter. In the second chapter, I describe a technique I developed for searching a genome for biologically relevant transcription factor binding site clusters. Using this methodology, we came to the surprising realization that many known Notch targets have statistically significant clusters of binding sites for Suppressor of Hairless [Su(H)], the transcription factor responsible for transducing the Notch signal. In chapter three, I validate a novel cluster of Su(H) binding sites, identified by my *in silico* study, residing within the *numb* gene. The investigation of this enhancer not only validates my bioinformatic approach, but also serves to illuminate several as of yet unappreciated aspects of bristle development. In chapter four, I take a different approach to finding new *cis*-regulatory sequences. By composing a hypothetical dual input code for neural precursors, we identify a cluster of binding sites that implements this code through a candidate gene approach. Finally, in chapter five, I present the discovery of an ancient transcription factor binding site, conserved for >700 million years. This finding establishes that although *cis*-regulatory change is a major engine for evolution, some transcriptional
linkages can withstand the constant erosion of sequence turnover for extremely long periods.
INTRODUCTION
The development of an organism is inextricably linked to the sequence of its genome. From the point of fertilization, each gene in the genome is constrained by individual control of transcriptional activity in both time and space. The genome sequence holds the key to both the proteins that are encoded by each gene, and the complex information needed to direct each gene’s own expression pattern. During the past decade, the sequences of many organisms have been elucidated, giving us, in principle, the key to their developmental programs. Through the mass of knowledge accumulated on gene and protein structure, similarity searches, EST databases, etc, we have developed extremely effective methods of comprehending the protein-coding fraction of the genome. What have lagged are techniques that leverage an interpretation of the non-protein coding portion of the genome. This massive undertaking requires predictive cognizance of a diverse array of mechanisms that regulate transcription, translation, mRNA degradation, alternate splicing, chromatin structure modification, and a host of mechanisms that have yet to be discovered. Through careful reductionist studies on individual non-coding regions, we are beginning to derive rules that we can attempt to apply to the whole genome.

In this thesis, I describe the tools, methodologies, and observations that I have developed while using whole genome sequences to study transcriptional regulation, and demonstrate how these techniques have
applied to my chosen biological model for development: the *Drosophila* bristle organ.

**Transcriptional regulation of gene expression**

All genes are subject to some form of transcriptional regulation: perhaps the gene is turned on constitutively, or in response to an extracellular stimulus. A gene’s transcription may be regulated to cause a burst of expression in only a few cells during a specific stage of development. These complex and diverse expression specificities are encoded in the genome of the organism by the sequence of DNA surrounding the gene. All genes contain a core promoter, which is required to recruit the basal transcriptional machinery, positioning the first base to be transcribed by the RNA polymerase complex (Smale and Kadonaga, 2003). Sequences outside of this ~50 base pair (bp) region respond to environmental cues of the cell to increase or decrease the firing rate of the polymerase complex, in turn modulating the amount of transcript produced. For this activity of enhancing basal transcription, these sequences are referred to as enhancer sequences. Enhancer regions may be positioned next to the gene, in introns, or even hundreds of kilobases away from the transcription start site. These sequences are generally portrayed as DNA stretches of 200-1000 bp, containing 10-20 binding sites for different DNA-binding transcription factors (Davidson, 2001).
The transcription factors that bind to specific sites within an enhancer convey on this unit different transcriptional properties. Some factors repress transcription while others may serve a role in activation. Many transcription factors act as both, showing context dependent activation or repression (Barolo and Posakony, 2002). Due to differences in binding specificities of individual transcription factors, the physical sequence of the DNA determines what factors can bind to any given region. The collection of transcription factors present in each nucleus then determines which of these factors are bound. The analogy has often been made that the DNA sequence of an enhancer is the blueprint, which the nucleus must read to determine if the gene is turned on or off. It is common for an enhancer region to drive only a subset of the gene’s total expression pattern: all of the enhancer sequences for the gene are required to completely recapitulate its pattern of expression. The development of an organism can be viewed at some level to represent the additive outcome of each gene’s complement of multiple expression “blueprints”.

The elusive enhancer element

Several obscuring properties of enhancer elements retard their a priori identification in whole genome sequences. The canonical enhancer model (a collection of 10-20 transcription factor binding sites) itself poses the problem
that we have very few transcription factors for which we understand in great
detail their sequence specific binding properties (Wray et al., 2003). This
problem is compounded by the relative size of elements with respect to the
average size of a gene: looking for a 500 bp element across 30 kb is much
akin to looking for a needle in a haystack. Furthermore, when sequences of
several enhancers that drive diverse specificities are compared, no one has
been able to find a common sequence property that distinguishes enhancer
sequence from sequences with no regulatory function (Wray et al., 2003). This
observation is troublesome when one also considers that it is generally difficult
to recapitulate an enhancer synthetically by multimerizing the functional
sequences for an enhancer artificially – these constructs are usually unable to
recapitulate the expression pattern of the module being imitated. Therefore,
we are confronted with a list of transcription factors that we don’t know too
much about, a vast collection of DNA sequences that we cannot discern a
function for, and a nebulous sequence component that is not conserved, but
seemingly required.

Two main strategies have been employed in the \textit{a priori} identification of
enhancer elements in whole genome sequences: clustering and conservation
of transcription factor binding sites. The motivation for attempting to discover
enhancers by clustering takes root in the simple observation that many
functional enhancer elements contain clusters of binding sites for the same
factor (Davidson, 2001). Using this property of many known enhancers, I, and several others have shown that novel cis-regulatory sequences can be identified in the genome (Berman et al., 2002; Markstein et al., 2002; Rebeiz et al., 2002) (Chapter 2). These studies represent an advance in understanding non-coding DNA based solely on the sequence of nucleotides.

The other most widely used tactic for identifying enhancer sequences *a priori* has been to search distantly related species for conserved non-coding DNA. This strategy is referred to as *phylogenetic footprinting*. Some methods look merely for the maintenance of large stretches of identity, as has been used in mammalian systems (Loots et al., 2000). However, this type of search criterion has not been successful in insects, and it appears that searches for the conservation of individual binding sites is a more practical technique for finding enhancers *a priori* (Berman et al., 2004). However, a defect in all of these techniques is that functional sequences need not be conserved, nor clustered. It is only through continued study of these critical regulatory regions, and the iterative application of new knowledge to genome sequences that we can tackle this problem of finding cis-regulatory DNAs in genome sequence.

**The role of regulatory DNA in cell fate specification**

As mentioned above, cell fate specification and cellular development are tightly married to the transcriptional regulation of gene expression. During
development, each cell type has derived from a progenitor that had a greater potential for cell fate diversity. The restriction of fate requires the cell to understand its environment: where is it located on the A/P axis? In which tissue does it reside? Should it divide or not? One of the greatest advances in developmental biology is the realization that cells use a limited number of evolutionarily conserved communication pathways, and that a component of each pathway’s mode of action is the same abstract outcome of a transcription factor modifying its regulatory activity. Therefore, to study cell fate, one must eventually confront transcriptional regulation of differential gene activity. One of the premier settings for studying cell fate specification and transcriptional regulation is during the development of external sensory bristles of *Drosophila melanogaster*.

**The *Drosophila* bristle model of cell fate specification**

The fruitfly is covered with hundreds of sensory structures that allow it to perceive its environment through the impermeable barrier of its chitinous cuticle. Although there are several subtypes of sensilla that decorate the adult *Drosophila*, the large (macrochaeta) and small (microchaeta) mechanosensory bristles on the notum of the fly (Fig. I.1A) are perhaps the best-defined class of sensory organs. Each bristle consists of multiple cells (Fig. I.1B), related by lineage (Hartenstein and Posakony, 1989). The shaft and socket cells secrete
external cuticular structures, and have large polyploid nuclei. The socket cell lies closest to the cuticle, and secretes a doughnut shaped cuticular lip that holds the bristle in place. The shaft cell secretes a filamentous bristle, dying once it has completed this task. The bristle organ contains a bi-polar neuron that senses the mechanical stimulus. Apically, the dendrite forms a junction with the sheath and socket cell to make up the sensory apparatus. Basally, the axon migrates to the CNS. The sheath cell wraps around the neuron to form a glial-like sheath structure (For a detailed review of bristle physiology, see (Keil, 1997)). A glial-like cell is also associated with notum microchaetes, but dies shortly after birth (Fichelson and Gho, 2003).

Both microchaetes and macrochaetes of the notum develop through an identical cell lineage (Fig. I.2B) (Gho et al., 1999; Reddy and Rodrigues, 1999). Each organ begins as a specialized epidermal cell, the sensory organ precursor (SOP). The SOP divides four times to give rise to all five cells of the organ (Fig. I.2B). First the SOP divides to generate a pIIa and pIIb cell. Next, the pIIb cell divides to create the pIIIb and the pIIIb-sib cell. The I Ib-sib cell exhibits several glial-like properties, yet migrates away from the organ, and undergoes apoptosis (Fichelson and Gho, 2003). The pIIa cell next divides to yield the socket and shaft cells. Finally the pIIIb cell division results in the sheath and neuron cells. A microchaete organ develops through all of these stages in a period of about 8 hours.
Figure I.1. Bristle organs of *Drosophila melanogaster*. (A) The adult thorax (notum) is covered with both large and small bristles, macrochaetes, and microchaetes respectively. (B) Both of these classes of notum bristle exhibit a stereotypical cellular composition. The shaft and socket cells have large polyploid nuclei, and secrete external cuticular structures responsible for receiving mechanical stimuli. A single bipolar neuron, encased by a sheath cell innervates the sensory scaffold created by the socket and shaft.

**Early development of bristle organs: specifying an SOP**

The founding cell of the bristle organ, the SOP, is selected from a group of equipotent epidermal cells, termed the “proneural cluster” (Cubas et al., 1991; Skeath and Carroll, 1991). The limits of the proneural cluster are defined by the expression of “proneural” basic-Helix-Loop-Helix (bHLH) transcriptional activators *achaete* and *scute*. In the absence of *achaete-scute* function, no bristles are formed (García-Bellido and Santamaria, 1978). Initially, proneural expression is uniform throughout the proneural cluster, but eventually, a single cell accumulates more *achaete-scute* expression, while levels in surrounding
cells are decreased (Cubas et al., 1991; Skeath and Carroll, 1991). The expression of *achaete* and *scute* is initiated by tissue-specific enhancers that integrate positional information to place each bristle in the appropriate location (Garcia-Garcia et al., 1999; Gomez-Skarmeta et al., 1995). Once a proneural cluster has been established, the Notch pathway plays an important role in the restriction of proneural gene expression to a single SOP, as well as many subsequent steps in bristle formation (Fig. 1.2A).

**Multiple roles for the Notch pathway in bristle development.**

The Notch pathway represents an ancient cell-cell communication system utilized throughout metazoan life (Lai, 2004). The Notch receptor is an integral membrane protein that binds to ligands Delta and Serrate (Fig. I.2A). Upon ligand binding, the Notch receptor is cleaved multiple times, releasing its intracellular domain ($N^{IC}$) into cytoplasm of the receiving cell. Once free, $N^{IC}$ enters the nucleus where it binds the transcription factor Suppressor of Hairless [Su(H)]. In cells that have not received a Notch signal, Su(H) is bound at target sites in complexes with co-repressors which prevent target genes from being transcribed. Once the Su(H)-co-repressor complex encounters $N^{IC}$, repression is relieved, and Su(H), $N^{IC}$, and the protein Mastermind form a transcriptional activation complex (Lai, 2004).
Mutations in components of the Notch pathway have effects on multiple phases of bristle development (Fig. 1.2B). This property is evidenced nicely by the phenotype of temperature-sensitive alleles of Notch (Hartenstein and Posakony, 1990) and Delta (Parks and Muskavitch, 1993). When incubated at the non-permissive temperature during lateral inhibition, resulting flies show supernumerary SOPs. In contrast, when shifted to the non-permissive temperature after SOP division, sister cell fates are disrupted such that only one fate is adopted by both cells. In both settings (SOP selection and the bristle lineage), the Notch pathway is thought to distinguish between two alternate fates: (1) epidermal cell vs SOP, (2) pIIa vs pIIb, (3) pIIlb vs pIIlb sib, (4) socket vs shaft, and (5) sheath vs neuron.

The role of Notch in bristle development is best understood in the context of SOP selection. In this setting, it has been demonstrated that signaling through the pathway is required to turn on genes of the Enhancer of split [E(spl)] complex in non-SOP cells of the proneural cluster (Bailey and Posakony, 1995). The bHLH-repressors (bHLH-R’s), and bearded family members of this gene complex are required to single out an SOP (Heitzler et al., 1996). However, the exact mechanism through which the SOP is initially selected from the cluster remains a mystery. The \( E(spl) \) bHLH-repressors are
Figure I.2. The Notch pathway during bristle development. (A) Summary of the Notch pathway. Depicted are two cells signaling to each other via the Notch signal cascade. When the transmembrane receptor Notch is bound by its ligand, Delta, the receptor is cleaved, releasing the intracellular domain (N\textsuperscript{IC}), from the cellular membrane. N\textsuperscript{IC} enters the nucleus, where it binds to the transcription factor Suppressor of Hairless [Su(H)], causing the dissociation of Su(H) from repressive complexes (“R”). N\textsuperscript{IC}/Su(H) then form a transcriptional activation complex. Thus, signaling through the pathway causes a switch from repression to activation.

(B) Multiple roles for Notch signaling during bristle development. In the proneural cluster, Notch signaling is used to ensure the specification of only a single sensory organ precursor (SOP). The non-SOP cells respond to signaling through the pathway. Once the SOP divides, Notch signaling is required to distinguish between sister fates of the daughter cells (pIIa and pIIb). This binary cell fate decision is used iteratively for all four sister cell pairs in the \textit{Drosophila} bristle lineage. pIIa, pIIb, sheath, and socket cells are Notch-responsive, while pIIb, pIIb sib, neuron and shaft cells are resistant to Notch signaling. The pIIb sib cell undergoes apoptosis (Red “X”).
thought to carry out the job of turning off genes that would normally be activated in the SOP, including the proneurals. This is evidenced by the finding that a VP-16/\(E(spl)\) bHLH-R chimera can cause overselection of SOPs, ultimately forming bristle tufts (Jiménez and Ish-Horowicz, 1997). In Chapter 4, I present the characterization of the first two enhancers to be placed downstream of \(E(spl)\) bHLH-repressors.

Contrary to the wealth of knowledge regarding proneural clusters, the mechanism by which Notch signaling distinguishes sister cell fates in the bristle lineage remains murky. However, during these sister cell fate decisions, the basis for Notch signaling asymmetry is fairly well known. The protein Numb accumulates in the SOP, and during division is asymmetrically segregated to one cell (the pIIb cell) (Rhyu et al., 1994). In the pIIb cell, Numb is required to antagonize Notch signaling. Thus, in a \(numb\) mutant, the pIIb cell is transformed to a pIIa cell. In all sister cell fate decisions of the bristle lineage, \(numb\) is used in a similar way to bias the decision mediated by Notch (Van De Bor and Giangrande, 2001; Wang et al., 1997).

The mechanism by which Notch signaling dictates cell fates in the bristle lineage remains a mystery. What Su(H) targets are activated in the bristle lineage? During lateral inhibition, there is clear evidence that the \(E(spl)\) complex genes are major effectors of Notch signaling (see above). Additionally, expression data has shown that some of these canonical targets
are present during lineage cell fate decisions (Nellesen et al., 1999). However, their importance has been uninvestigated, due to the multigenic nature of the \( E(spl) \) complex, and the confounding effects that \( E(spl) \) mutations have on lateral inhibition. The only other documented Notch target in the bristle lineage is Su(H) itself (Barolo et al., 2000). A regulatory element downstream of the Su(H) transcription unit contains seven high affinity Su(H) binding sites, and drives strong expression of Su(H) in the socket, a Notch responsive cell. Interestingly, inactivation of this enhancer does not effect the specification of the socket cell fate, but instead leads to differentiative defects of the socket, which incapacitates the mechanosensory function of these organs. In Chapter 2, we use a bioinformatic approach to find new Notch targets by looking for statistically significant clusters of Su(H) binding sites. The goal of finding new targets of Su(H) in the bristle lineage is realized in Chapter 3, where one of these clusters is characterized, and found to be a Notch responsive enhancer element of the \textit{numb} gene. A detailed comparison and functional analysis of these bristle lineage enhancers may be the key to understanding how the Notch pathway can be deployed repetitively to stimulate distinct cell fates.

**References**


Wang, S., Younger-Shepherd, S., Jan, L. Y., and Jan, Y. N. (1997). Only a subset of the binary cell fate decisions mediated by Numb/Notch signaling in

CHAPTER 1
GENEPALETTE: A UNIVERSAL SOFTWARE TOOL FOR GENOME
SEQUENCE VISUALIZATION AND ANALYSIS
Abstract

To make effective use of the growing host of complete genome sequences, biologists must have easy-to-use software tools that allow them to visualize, analyze, and modify genome data in an interactive and generalized manner. In an effort to bridge the gap between genome and researcher, we have created GenePalette (www.genepalette.org), a desktop application that can access any genome sequence and display the positions of various features [e.g., transcription factor binding sites (TFBSs)] relative to the introns and exons of annotated genes. Written in Java, GenePalette can run on all Java-supporting operating systems (Mac, PC, Unix, Linux). Annotated sequence encompassing the majority of public genome data is rapidly retrieved from GenBank or Ensembl. The software provides intuitive access to the selected genomic region through three interface components: a colorful graphical display showing a schematic of genes and features; an annotated sequence view in which features and genes are highlighted directly on the sequence; and the selectable raw sequence. The three interface components are fully integrated and presented on one page, permitting the user to move easily between representations at different levels of resolution, ranging from kilobases to individual nucleotides. GenePalette is a particularly powerful platform for analyzing the organization of cis-regulatory elements and designing wet-lab experiments to investigate them.

© 2004 Elsevier Inc. All rights reserved.

Keywords: Desktop application; GenBank; Sequence retrieval; Genome sequence; Genome annotation; Sequence analysis; cis-regulatory; Transcription factor; Binding sites

Introduction

The purpose of complete genome sequencing is to accelerate understanding and experimental discovery in biology and biomedicine. Ideally, these powerful sequence resources should be placed as rapidly as possible into the hands of all biologists (Collins et al., 2003). Achieving this goal requires publicly available tools that allow genome data to be accessed in a way that requires little or no computational knowledge. Furthermore, these access tools should lend themselves to the efficient design of "wet-lab" experiments.

Several genome browsers exist as web-based environments (Birney et al., 2004; Karolchik et al., 2003), yet there is the sentiment that these static web interfaces do not provide a sufficiently rich interactive user experience (Lorraine and Helt, 2002). The web interface is an indispensable data source for housing and presenting the latest information about a genomic region, but lacks the key components required for effective experimental design. Frequently, one wants to view the positions of user-defined features [e.g., transcription factor binding sites (TFBSs)] in the context of gene annotations. It is also common to want to design primers or a cloning strategy based on the positions of these features. When an experiment is designed in this way, it is crucial to be able to revisit the region of interest to ask new questions about the existing design or about additional features.

As an alternative to web interfaces, desktop application tools have several advantages. Users can save their work and return to the saved state. Files saved from a properly designed universal application can be shared between collaborators to communicate the most recent view of a project's progress. It is simple to implement desktop applications without having to create permanent and complicated server-client interactions. As noted above, applications
allow a greater depth of user interaction, such that simple operations yield immediate results.

We have designed a universal desktop application called GenePalette that crosses both platform and organism boundaries to bring feature-oriented genome sequence manipulation to all flavors of biologist. Implemented in Java, GenePalette has been successfully ported to all major operating systems (Mac, PC, Linux, Unix). The program can be freely downloaded by academic and nonprofit users through our website (www.genepalette.org), and is installed by a user-friendly installation program. Via access to GenBank, GenePalette can retrieve sequence and annotation from any deposited genome. Through its intuitive, single-window interface, GenePalette provides the core set of sequence analysis, experimental design, and graphical output functions needed by biologists to frame questions about a given genome segment. It is a particularly effective and efficient tool for examining the disposition of cis-regulatory elements in genomic DNA, and for planning and modifying experiments based on the results.

Materials and methods

GenePalette was written with the Java JDK version 1.3.1 on a Macintosh G4 running OS X version 10.2.6. The software was packaged into platform-specific installers using InstallAnywhere 5.5.1 Now! (Zero G).

The software is freely available to academic and nonprofit users through the website http://www.genepalette.org. For new users, the website provides access to both a Quick Reference guide and a more detailed User’s Manual, both in Portable Document Format. Commercial users may go to http://invent.ucsd.edu/ for information on purchasing a license for GenePalette.

Results

Sequence access via GenBank

When designing a data import system for GenePalette, we wanted the software to support the genome sequences of as many organisms as possible, to cater to the broadest array of biological scientists. Moreover, we envisioned that any sequence access capability should be easy to use—one should not have to remember a genomic location, or go through too many steps to load the sequence of the desired genomic region. GenBank (Benson et al., 2004) provides a simple and relatively uniform interface to genome sequence annotation data through the GenBank Flat File. GenBank holds nearly all of the sequenced genomes available in the public arena, including human, mouse, fly, mosquito, worm, Arabidopsis, and over a hundred microbial genomes. Within a genome sequence’s GenBank Flat File, information about gene annotation is stored in a uniform manner. Sequences in GenBank can be accessed quickly using the EUutils suite of programs to download dynamically only the subregion of sequence desired. An added benefit of GenBank as a source of data is that all GenBank records are indexed in the Entrez system (Wheelier et al., 2004) by gene name, such that an Entrez query containing a gene symbol will locate genomic regions containing the given gene.

To access a genome sequence segment in GenePalette, the user completes three simple interface-prompted steps (Fig. 1). (1) To initiate the data-access process, the user types an Entrez query (in most cases a gene symbol) into a dialog box. The query is then submitted to the Entrez web server, and a list of sequence records that match the query are returned. (2) Next, the user selects a sequence of interest, usually a genomic DNA contig, and the genes annotated in the sequence’s GenBank record are then loaded into a gene selection dialog. (3) The user is subsequently prompted to select genes to import, defining the segment that will be downloaded. When all three steps are complete, the desired sequence segment is retrieved from GenBank using the Entrez EUutils package.

As an alternative to sequence access via GenBank, GenePalette supports the import of GenBank Flat Files from external sources. Many programs and websites, including Ensembl and the NCBI MapViewer, can export annotation data in the form of a GenBank Flat File. Using this option, the user pastes a flat file from the external source into a dialog box to load the external data.

The GenePalette interface

When a user is inspecting and manipulating a genome sequence segment, there are several desirable perspectives. At the large scale, the user will want to see how her gene or feature fits into the overall landscape of surrounding genes. At this scale, it should be straightforward to measure distances and to access features of interest. Furthermore, from this type of view, it should be easy to move to the other views provided by the interface. At a smaller scale, it is important to be able to see how individual genes and features are disposed on the sequence at single-nucleotide resolution. This scale would allow one to view the sequence surrounding a TFBS, for example, or to inspect an exon splice junction. Finally, for the purpose of input to other applications, the capacity to view and select the sequence itself is essential.

In the GenePalette interface, we have combined these three perspectives to create a highly integrated single-window environment (Fig. 2). Multiple loaded sequences are managed independently in their own windows, such that there is no confusion as to which window belongs to what sequence. At the top of the window is a selectable version of the raw sequence (“Sequence View”; Fig. 2). Within this component, information about the current selection (length and sequence coordinates) is reported. Beneath the Sequence View is a “Markup View” where both genes and user-defined features are highlighted directly on the double-stranded sequence. In
Fig. 1. Accessing sequence data in GenBank using GenePalette. The user enters an Entrez keyword query (e.g., a gene symbol) into GenePalette (Step 1), the query is sent to the Entrez system at NCBI, and the results are retrieved (center left). The desired sequence record is selected (Step 2, “Select” check box), and the annotated genes associated with that sequence segment are loaded and presented in a table (center right) which allows the user to choose which gene sequences to import into GenePalette (Step 3, “Include” check boxes). Optionally, the user can modify the amounts of upstream and downstream flanking sequence to be loaded (Step 4, the default values are the full distances to the adjacent genes).

In this display, features appear as boxes enclosing the relevant sequence (Fig. 2), and exons are displayed by color-coding the corresponding bases (Fig. 3, Markup View, left). Below the Markup View is a component that displays the genome segment schematically (Fig. 2, Graphical View). Finally, at the bottom of the main window are panels containing tables of data pertaining to annotated genes (left) and added sequence features (right). The table panel for each transcript or feature is accessed by a selectable drop-box (Fig. 2).

The major visual component of the GenePalette window is the Graphical View. The scale of this view ranges from 1 bp per pixel to 100 bp per pixel and is controlled instantaneously by a slider button (compare the scale of Graphical Views in Figs. 2 and 3). Sequence regions that lie outside the Graphical View panel at the current scale may be rapidly browsed or accessed via a scroll bar at the base of the panel (Figs. 2 and 3). The Graphical View depicts genes as connected boxes (illustrating exons and splicing patterns), with an arrow at the transcription start site to indicate the direction of transcription. Overlapping genes or transcript variants are arrayed vertically as seen in Figs. 2 and 3. User-defined features can be represented in the Graphical View either as colored shapes or as textual symbols that appear on the end of short lines perpendicular to the line that symbolizes the DNA. Features above the DNA line are on the top strand; features below are on the bottom strand (Figs. 2 and 3).

The graphical view is integrated with all components of the interface.

The schematic representation in the Graphical View serves as a nexus for accessing the finer details of a genome.
Fig. 2. The GenePalette interface provides three perspectives of a genome sequence segment in a single window. The Sequence View (top; see brackets at right) allows the user to select and copy the entire or a subset of the raw sequence. In the Markup View, genes and user-defined features are highlighted directly on the sequence, providing explicit information about the arrangement of these elements at single-nucleotide resolution. The Graphical View presents a detailed diagrammatic representation of a sequence region, showing both genes and added sequence features. Separate tables at the bottom of the window (Data View) contain information, respectively, about genes (including transcript variants; Transcript Data) and features (Feature Data), and also permit the user to control the appearance of these elements in the Graphical View.

Segment. Several operations performed with the mouse in this display will elicit a response from the other components of the interface (Fig. 3). Clicking an exon in the Graphical View will cause the sequence to be selected in the Sequence View (Fig. 3, solid blue arrow and blue-shaded bases). This selection is convenient because it reports the exact details of the selected sequence, including length, and allows the user to copy the exon’s sequence rapidly. Clicking an exon also brings up the table for the transcript containing the exon in the Transcript Data panel (Fig. 3, dashed blue arrow), highlighting the row for that exon in the table. Finally, clicking on an exon causes the exon to be outlined in red in the Graphical View (Fig. 3).

Another action available from the Graphical View is making a graphical selection by dragging a box over part of the sequence schematic. When a box is dragged in the Graphical View (Fig. 3, black outline), the sequence corresponding to the boxed region is selected in the Sequence View (Fig. 3, solid red arrow), and the length and other information about this selection is updated. Another effect of box-dragging is that the sequence region contained in the box is displayed in the Markup View (Fig. 3, dashed red arrow). Box-dragging is a useful operation for quickly determining the distance (shown above the Sequence View) between two elements in the Graphical
View, for viewing the detailed context of a group of features in the sequence (Markup View), or for copying a sequence stretch containing interesting features for input into other programs.

A third action performed from the Graphical View is clicking on a feature (Fig. 3, green arrow). When a feature is clicked, a vertical red marker arrow is positioned directly underneath the clicked feature in the Graphical View. The click also invokes the production of a Markup View (Fig. 3, solid green arrow), highlighting the clicked feature and any other features 50 bp upstream and downstream of the feature. Lastly, a feature click causes the table for that feature to be visible in the Feature Data panel (Fig. 3, dashed green arrow), and the row for that feature to be highlighted. This operation is thus extremely useful for inspecting individual features and their context in the sequence, and for gaining access to the components that modify a feature’s appearance (Feature Data panel).
Connections between other components of the interface

Although the Graphical View is the most interconnected view, every component of a GenePalette window elicits a response in one or more of the other components of the interface (Fig. 4). Selections in the Sequence View are depicted schematically by a box in the Graphical View. Each base in the Markup View can be clicked, causing that base to be selected in the Sequence View, and also repositioning the red marker arrow in the Graphical View. As individual exons or features in the Transcript/Feature Data tables are clicked, their positions are represented schematically in the Graphical View. This high level of connectivity improves the overall efficiency of inspecting a sequence segment and designing experiments based on its current annotation.

Searching for and displaying specific features in the sequence

A key motivation for the design of GenePalette was to give investigators the ability to interact more deeply with a genome segment through the visualization of personalized features within the sequence. A major need for many researchers is to know where predicted TFBSs lie in the region around their favorite genes; others are constantly in need of an easy way to design primer pairs to amplify genomic DNA segments; still others must routinely visualize the locations of single-nucleotide polymorphisms (SNPs) in genomic regions of interest.

The predominant type of feature search performed in GenePalette makes use of consensus sequences written in the IUPAC code. The sequence(s) defining the consensus and the name of the feature are input into a dialog box, and all matches found in the sequence are reported both in the Graphical View and in a Feature Data table (Figs. 2 and 3).

For example, the Graphical Views in Figs. 2 and 3 display the location of high-affinity binding sites for the Drosophila transcription factor Suppressor of Hairless [Saf(1)] in the region surrounding the Suppressor of Hairless gene (Barolo et al., 2000).

For any application of the feature search capability, it may be highly desirable to store large numbers of consensus sequences for repeated searching on a routine basis. GenePalette supports stored feature libraries for this purpose. Multiple feature libraries can be loaded at the same time, and GenePalette provides menu-driven support for manipulating them. In keeping with the Java philosophy, libraries are completely portable from one platform to another. GenePalette comes equipped with a standard library of approximately 200 restriction enzyme sites.

Sequence manipulation and modification features of GenePalette

In addition to the core graphical interface and feature searching functions, GenePalette supports many menu-driven operations to aid in the manipulation and modification of a loaded sequence: (1) sequences can be trimmed and reverse complemented; (2) menu options are provided for extracting spliced cDNA sequences from an annotated genomic DNA segment containing one or more transcripts; (3) alternatively, selected sequence regions can be masked with N’s based on their coding or noncoding content; (4) annotated transcripts can be modified in a special editor dialog box, permitting the user to incorporate new, unpublished data on gene structure (transcription start sites, intron boundaries, etc.). Collectively, these options expand the convenience and versatility of sequence access and visualization in GenePalette.

Local access to stored sequence collections

Although remote access to GenBank through Entrez is dynamic, and gives access to the most up-to-date sequence and annotation information, there are several benefits associated with accessing a genome sequence locally. Not only does it save the time needed to download information from GenBank, it also gives the user nearly complete access to a genome sequence when working off-line. GenePalette implements an effective system for local access, providing pre-assembled whole genome collections for many model organisms at the GenePalette website (www.genepalette.org). Within the program, the user can make his own collections by interactively downloading multiple sequences through Entrez.

Once a collection of sequences is established on the user’s system, the sequence can be accessed in two ways. First, local sequences are transparently accessed in the normal Entrez query process described above (see Fig. 1). When a query is entered, it is sent to the Entrez server at the National Library of Medicine. Query matches are returned.
in the normal fashion, and the list of matching sequence entries is in turn matched against an index of local sequences. If a sequence is available locally, this information is conveyed in the second column of the sequence selection dialog ("Source", Fig. 1), which changes from "Remote" to "Local". If a "Local" sequence is selected by the user, the local copy is accessed to retrieve the list of contained genes, instead of the sequence being downloaded. After a sequence segment is selected, the sequence is loaded from disk into GenePalette. Under this system, the only internet usage by the software is the initial Entrez query, allowing the user to take full advantage of the search capabilities of the Entrez system.

Just as importantly, a local sequence collection permits access to sequence data in a manner that is completely independent of an internet connection (e.g., while on an airplane). Under this system, the user can select any symbol for the gene she wants to browse, as it appears in GenBank. GenePalette first indexes all genes contained in the local sequence collection and then cross-references the desired symbol against this index. When the symbol is found, the user can select genes from the local copy of the data record that contains the symbol. Although this method does require some knowledge of gene nomenclature, it gives the fullest possible access to sequence data without connecting to a remote database.

Image output options

When investigating a particular gene or genome segment, it is often desirable to create schematic images of the region, drawn to scale, for the purposes of presentation or publication. Many programs have the ability to export graphical outputs, but it is often the case that these outputs are static images that nevertheless require further doctoring to achieve presentation quality. In addition to static GIF outputs of the Graphical and Markup Views, GenePalette can also export the contents of these views in Postscript format. The Postscript outputs can be manipulated in graphical packages such as Adobe Illustrator, such that every element of the output can be modified by the user. Thus, GenePalette’s graphical output capabilities are versatile, permitting the user to create customized printouts, posters, presentations, and even journal figures.

Discussion

A user-friendly platform for investigating cis-regulatory elements

The guiding motivation in our development of GenePalette was to make feature-oriented genome sequence visualization accessible to all biologists. To create a truly interactive environment, it was decided that a desktop application would provide the closest and most responsive interface between sequence and user. We selected Java to make GenePalette functional on all of the most common computing platforms. Combining these ambitions, we have created an easily installed desktop tool that gives access to all publicly available genome sequences through an intuitive user interface. The applicability of the software to any genome allows users to investigate genomic regions from diverse species at an unprecedented level of continuity. No longer must users learn the in’s and out’s of three different genome browsers to study intron placement or the positions of TFBSs in *C. elegans*, *M. musculus*, and *D. melanogaster*.

Although GenePalette has many functions useful for general genome browsing, the principal rationale underlying the software was the creation of a platform for the rapid and efficient investigation of cis-regulatory linkages. To do this, GenePalette supports user-defined feature searches and the storage of feature lists in libraries. At this time, GenePalette is not packaged with a database of TFBSs, yet with the portability of GenePalette libraries, we envision that our website could serve as a central resource to mediate the sharing of TFBS consensus sequences between participating members of the community. For many purposes, a binding specificity defined by an expert investigator is preferable to one taken from a TFBS data warehouse such as TRANSFAC (Matys et al., 2003). A GenePalette library composed of such expertly assembled binding site data would represent a very valuable general resource, and we encourage users to contribute.

Sequence access based on GenBank

When selecting a data source for GenePalette, we focused on GenBank because of its searchable interface and broad collection of genome sequences. To cover as many applications of the software as possible, we have also incorporated a system for importing external GenBank Flat Files, adding the capacity to use annotations from several other data sources, including Ensembl (Birney et al., 2004). The Distributed Annotation System (DAS) (Doxell et al., 2001) also provides a generalized method of access to genome annotation data. The major advantage of DAS is its controlled vocabulary, ensuring that the syntax will always be compatible with a given annotation parser. However, this strength is outweighed by the inability to search a DAS server by gene symbol. To make effective use of a DAS browser, one must either know the exact location of the gene, or have a front-end application that searches an external database. As DAS and other general annotated sequence resources evolve, GenePalette will be modified to make use of them.

Future directions for GenePalette

As genome sequence data accumulate and annotations mature, we plan to continue to develop the GenePalette software as a feature-driven sequence manipulation and
visualization tool. Planned improvements include local alignments, enhanced feature support, and augmented interface connectivity. To broaden GenePalette’s feature visualization capability, it would be extremely useful to be able to perform local sequence alignments, interactively viewing the positions of the aligned regions within the Graphical View. This type of operation is useful for the localization of features that do not match the DNA sequence perfectly, or are too large to be usefully treated as a generic feature. Local alignment support would also aid in the evaluation of gene annotation, though gene prediction is not a major purpose of the software. Several improvements in the way GenePalette handles user-defined features are contemplated. First, as new ways to depict and analyze transcription factor binding specificities arise, we plan to incorporate these conventions into the software. There are other types of sequences currently treated as generic features that could also benefit from having more specific definitions within the GenePalette data structure. Primer pairs are easy to denote as simple features, yet it would be useful to treat them as a distinct entity, to convey information about base content and product size. Restriction site representation would also be improved by more a defined data specification, including the position of the cut site and information about methylation site overlaps.

There are several developments on the horizon that promise to change the way genomes are browsed. As genome sequences accumulate, the need for uniform data representation becomes critical. In GenBank, gene annotation has been consistent from genome to genome, yet there are many other features that could benefit from such consistency, including SNPs, repeats, microRNAs, transposable elements, and tRNAs. The uniform annotation of these features would allow programs like GenePalette to deal with these elements in a way that fits both the nature of the feature and the needs of end-users for incorporating this information into their analysis.

A major benefit of fully sequenced genomes is the great potential to explore evolutionarily conserved features through comparative genomics. Although tools exist to assess the conservation of TFBSs (Loots et al., 2002), for example, these functions would be most useful in an environment where the user can inspect the context of the sequence element, and easily design experiments based on the results. As a universal language for genome annotation and comparative analysis develops, we hope to bring these fairly complex tasks to GenePalette while maintaining the simplicity and user-friendliness of the interface.

For the present, however, we offer GenePalette as a much-needed bridge between the genomes of a wide variety of organisms and the essential reductionist studies by which we will find out just how those genomes work.

Acknowledgments

We thank Scott Barolo and Par Towb for discussions about the design of the interface. Thanks to members of the Posakony lab, and the lab of Bill McGinnis, for testing the software. This work was funded by NIH Grant GM062279 to J.W.P.

References


Chapter One, in full, is a reprint of the following publication. Rebeiz M, Posakony JW. GenePalette: a universal software tool for genome sequence visualization and analysis. Dev Biol. 2004 Jul 15;271(2):431-8. I was the primary researcher and author and James W. Posakony directed and supervised the research that forms the basis for this chapter.
CHAPTER 2
SCORE: A COMPUTATIONAL APPROACH TO THE IDENTIFICATION OF \textit{CIS-}
REGULATORY MODULES AND TARGET GENES IN WHOLE-GENOME
SEQUENCE DATA
SCORE: A computational approach to the identification of cis-regulatory modules and target genes in whole-genome sequence data

Mark Rebeiz, Nick L. Reeves, and James W. Posakony*

Division of Biological Sciences, Cell and Developmental Biology, University of California at San Diego, La Jolla, CA 92093-0349

Communicated by Michael S. Levine, University of California, Berkeley, CA, May 23, 2002 (received for review February 7, 2002)

A large fraction of the information content of metazoan genomes resides in the transcriptional and posttranscriptional cis-regulatory elements that collectively provide the blueprint for using the protein-coding capacity of the DNA, thus guiding the development and physiology of the entire organism. As successive whole-genome sequencing projects—including those of mice and humans—are completed, we have full access to the regulatory genome of yet another species. But our ability to decipher the cis-regulatory code, and hence to link genes into regulatory networks on a global scale, is currently very limited. Here we describe SCORE (Site Clustering Over Random Expectation), a computational method for identifying transcriptional cis-regulatory modules based on the fact that they often contain, in statistically improbable concentrations, multiple binding sites for the same transcription factor. We have carried out a Drosophila genomewide inventory of predicted binding sites for the Notch-regulated transcription factor Suppressor of Hairless (Su(H)) and found that the fly genome contains highly nonrandom clusterings of Su(H) sites over a broad range of sequence intervals. We found that the most statistically significant clusters are very heavily enriched in both known and logical targets of Su(H) binding and regulation.

The utility of the SCORE approach was validated by in vivo experiments showing that proper expression of the novel gene Him in adult muscle precursor cells depends both on Su(H) gene activity and sequences that include a previously unstudied cluster of four Su(H) sites, indicating that Him is a likely direct target of Su(H).

Realizing the full promise of whole-genome sequencing projects depends on our ability to read and understand the tremendous informational richness contained therein. Computational methods for predicting novel protein coding genes in whole-genome sequence data are quite advanced, and various strategies for recognizing transcription units that generate noncoding RNAs are also available. However, these techniques address only part of the informational content of the genome. The complex blueprint that controls the utilization of the coding information in DNA is contained in the huge number of cis-regulatory elements, both transcriptional and posttranscriptional, that surround and invade the transcribed part of the genome. But it is clear that we are in our infancy in learning how to read the regulatory genome and thus decipher the blueprint.

Here we describe SCORE (Site Clustering Over Random Expectation), a computational method for identifying potential cis-regulatory modules and the target genes they serve. Transcriptional enhancer elements are generally quite compact, and they frequently include closely spaced binding sites for the same or multiple transcription factors. SCORE is designed to detect and statistically evaluate these structural features in whole-genome sequence data, and thus to reveal previously unrecognized enhancers. A conceptually similar method has been described recently by Markstein et al. (2).

Suppressor of Hairless (Su(H)) is the key transducing transcription factor for the Notch cell–cell signaling pathway (3–5), which is involved in a large variety of cell fate specification and patterning events during bilaterian development (6). The sequence specificity of DNA binding by Su(H) is well defined (4, 7, 8), and multiple direct targets of regulation by this factor and the Notch pathway have been identified (4, 8–11). That defined cis-regulatory modules associated with these targets frequently include multiple high-affinity binding sites for Su(H) (8–10) suggested that this factor might be favorable for evaluating the SCORE technique.

Materials and Methods

Whole-Genome Inventory of Consensus Sequence Matches. Perl scripts (available on request) were written to search the Drosophila genome for matches to binding site consensus sequences. Release 2 (October, 2000) chromosome arm sequences and gene annotation data were downloaded from the Berkeley Drosophila Genome Project web site (http://www.flybase.org). The positions of consensus sequence matches relative to known or predicted genes are calculated by the script, and the identity of the gene nearest the match is reported.

Clustering Analysis. Binding site clustering was assessed by tallying the number of additional sites lying to the right of each binding site, within a specified range of window sizes. Each successive binding site along the sequence was similarly treated as the left end of a new cluster set, and the tallying process was repeated. This method results in an inventory consisting of overlapping, but unique, clusters.

Monte Carlo Simulations. In anticipation of applying SCORE to clusters of binding sites for multiple factors, we chose Monte Carlo simulations as the means of estimating random clustering probabilities. We have verified that our simulations of single-factor binding site clustering conform to the Poisson distribution, as expected. Data sets of random site positions were generated by scattering a fixed number of sites (equal to the total number in the genome) randomly in a space equal to the size of the genome. One thousand of these randomized inventories were then analyzed by using the clustering algorithm just described, to estimate the probability P that a given cluster frequency observed in the genome (or greater) could arise by chance. Purity values for each cluster bin, expressed as percentages, were calculated as the observed frequency minus the mean random frequency, divided by the observed frequency. To classify bins as pure or enriched we used cutoffs of P < 0.005 and purities of >99% and >50%, respectively.

Plasmid Construction. Enhancer/promoter reporter constructs expressing nuclear green fluorescent protein were prepared by using the pShigE transformation vector (12). The 3′ terminus of both fragments representing the Him gene (2.2 and 4.0 kb) was nucleotide just 5′ to the translation initiation codon; both constructs thus included the Him promoter and the entire 5′ untranslated region.

Abbreviations: SCORE, Site Clustering Over Random Expectation; Su(H), Suppressor of Hairless; bHLH, basic helix-loop–helix.

*To whom reprint requests should be addressed. E-mail: posakony@ucsd.edu.
Fly Genotype. Su(H)\(^{399}\)/Su(H)\(^{589}\) was used as a Su(H) null genotype.

In Situ Hybridization. Digoxygenin-labeled antisense RNA probes representing the \(H\) gene were transcribed from a 2.4-kb genomic DNA fragment that contains the full extent of the predicted gene (13). \(In situ\) hybridization was performed as described (14).

Results

Global Inventory of High-Affinity Su(H) Binding Sites in the Drosophila Genome. The first step in our computational approach to identifying novel targets of Su(H) and the Notch pathway was to search the complete \(Drosophila\) genome sequence (13) for matches to a consensus definition (YGTGDGAA) of high-affinity binding sites for the Su(H) protein. This consensus was derived from a combination of known target sites in \(Drosophila\) and information from a random binding site selection analysis using the mouse ortholog of Su(H) (4, 7, 9). After eliminating the lower-affinity sequence TGGTGAGAA, the net consensus consisted of the five octamers CGTGGGAA, CGTGGAGAA, CGTGTTGAA, TGGTGAGAA, and TGGTGGAGAA. A total of 15,659 perfect matches to these five sequences was found in the global genome inventory, compared to the statistical expectation (based on mononucleotide frequencies) of 16,989. Thus, the real genome has a substantial (~8%) deficit of these high-affinity Su(H) binding site octamers. Of the observed 15,659 total sites, 9,886 were categorized as occurring in intergenic regions, 2,078 in predicted exons, 429 in putative 5\' and 3\' untranslated regions, and 3,266 in predicted introns. After removal of putatively exonic consensus matches, it was found that 5,592 unique genes are associated with position with at least one predicted Su(H) binding site.

Inventory of Su(H) Binding Site Clusters. Transcriptional enhancer modules that are responsive to Notch signaling activity frequently are characterized by the presence of multiple high-affinity Su(H) binding sites in a relatively small sequence interval (~1 kb) (4, 8–11). This finding suggests the utility of identifying unusual Su(H) binding site concentrations in the genome as a means of recognizing possible novel target enhancers. For this purpose we developed the cluster detection algorithm described in Materials and Methods. The search window size was varied by increments of 100 bp over the range of 100 to 5,000 bp to cover a large variety of enhancer module sizes and complexities. The complete matrix of Su(H) binding site cluster frequencies over this sequence interval range is presented in Fig. 1. We will refer to each position in the matrix as a cluster bin, wherein the frequency (number of a particular size of cluster \(x\) number of sites in \(y\) bp) is recorded.

Monte Carlo Simulation of Random Site Clustering. Any sufficiently large collection of sequence features will exhibit a certain degree of clustering even if they are distributed in the genome randomly. We used a Monte Carlo approach to simulate this background of randomly occurring Su(H) binding site clusters, to evaluate the statistical significance of the cluster frequencies observed in the real genome. One round of simulation is carried out as follows. A number of sites equal to the number of putative Su(H) sites in the \(Drosophila\) genome (15,659) is randomly positioned in a genome-space of the same size. Clustering of these site positions is then inventoried by the same algorithm used for cluster analysis of real genome data. Cluster frequencies for a large number of such random simulations are accumulated. These data are analyzed to determine for each cluster bin what fraction of random data sets show a cluster frequency equal to or greater than the frequency observed in the real genome. This process yields an estimated probability \(P\) that the observed real-genome cluster frequency is caused by chance.

When the cluster frequencies for Su(H) binding sites in the \(Drosophila\) genome were compared with those obtained in 1,000 random simulations, a large domain of bins with \(P = 0\) was observed (Fig. 1), revealing that the genome has a surprisingly high degree of extremely improbable clustering of Su(H) sites. This broad \(P = 0\) “valley” is flanked on both left and right by very steep “walls” rising to domains of bins with much higher \(P\) values (Fig. 1). On the left (Fig. 1), at all window sizes, the frequencies of clusters containing one site were reproduced in a high proportion of random simulations. Similarly, at the larger window sizes (>2,000 bp), the frequencies of clusters containing two sites were also reproduced regularly in random data sets (Fig. 1). On the right end of the frequency matrix (Fig. 1), there is an abrupt transition from very low \(P\) values to values of \(P \approx 0.1\). The latter represent cluster bins with a frequency of 0 in the real genome (Fig. 1) and 0 in the random data, resulting in \(P = 1\).

We further confirmed the extreme statistical unlikelihood of the Su(H) binding site clustering in the fly genome by a second method: carrying out probability analysis on randomly generated genome sequences. Using the mononucleotide base frequencies in the total fly genome sequence \((A = T = 0.288; C = G = 0.212)\), we constructed 50 random sequence versions of the genome that are identical in size to the real genome. Each of these random genomes was inventoried to determine its total number of Su(H) sites, and this fixed number was then used in 1,000 random clustering simulations by the Monte Carlo method described above, to obtain \(P\) values for each cluster bin in each random genome. We found that none of the 50 random genomes displayed probability plots (see Fig. 4, which is published as supporting information on the PNAS web site, www.pnas.org) even remotely resembling that of the real genome (Figs. 1 and 4A). As expected, we did observe occasional Su(H) binding site densities in the random genomes with values of \(P < 0.005\) (Fig. 4C). However, probability plots like the one shown in Fig. 4B were far more typical. We conclude that the \(Drosophila\) genome exhibits highly nonrandom concentrations of putative high-affinity Su(H) binding sites over a broad range of sequence intervals (Figs. 1 and 4A).

Mining Pure Bins of Su(H) Binding Site Clusters. When considering how best to make use of global data on clustering of transcription factor binding sites or other cis-regulatory elements in the genome, it is perhaps useful to distinguish two special categories of bins in the cluster frequency matrix. The first of these we term “pure” bins, defined as those with a nonzero frequency in the real genome data, but with a zero or very near-zero frequency in the random data sets resulting from the Monte Carlo simulations. These bins represent site densities that are so unlikely to have arisen by chance that they may be considered pure—i.e., statistically unexpected features of the real genome, uncontaminated by random clusters. A second and much more common category consists of “enriched” bins—those in which the expected random frequency is substantial, but in which the cluster frequency for the real genome is nonetheless considerably higher.

To quantify the differences among bins in this regard, we used a parameter we refer to as purity. We first computed the mean frequency of randomly expected clusters in each bin from 1,000 Monte Carlo permutations. Then, for each bin, we took the difference between the real-genome frequency and the average random frequency, to measure the excess number of clusters of a given size in the real genome over that expected by chance. This nonrandom excess value was divided by the real-genome frequency in the bin, yielding an estimate of purity (expressed as a percentage). Fig. 2 shows a plot of purity across the whole cluster matrix.

We first examined the contents of bins that met the criteria of having a purity of ≥99% and a probability value of \(P < 0.005\). Only 10 distinct genomic regions are identified by these bins. The first bin in the matrix where each region was found is marked in Fig. 2; the specifications of each region, including the identity of the nearest gene, are given in Table 1. We found that this short list is very heavily biased toward both known and logical targets of Su(H).
binding and regulation. Of the 20 genes, five Su(H), singleminded, Ocho, E(spl)m3, and E(spl)m2/E(spl)m3 have been shown previously by various experimental criteria (see Table 1) to be subject to transcriptional activation by Su(H) (9, 10, 15–19). Like E(spl)m3, two other genes (Hey and deadpan) encode basic helix-loop-helix (bHLH) repressor proteins. Most of these genes of this class in Drosophila are directly regulated by Su(H) (4, 8, 11, 20); moreover, the mouse Hey1 gene has already been shown to be directly activated by the mouse ortholog of Su(H) (21). Interestingly, Delta, which encodes a major ligand for the Notch receptor, also appears on this high-purity list, although a direct role for Su(H) in regulating Delta expression has not previously been suggested. Finally, we found that one major concentration of Su(H) sites contributes to many bins and is responsible for the large expanse of 100% pure, 5 regions of the cluster matrix. This highly unusual genome segment, which we call the ‘A Lot of Su(H) Sites (ALS)’ region, contains 25 predicted high-affinity Su(H) sites within 5.3 kb. Overall, it is clear that the very high-purity, low-probability bins are extremely selective detectors of bona fide Su(H) targets.

**Mining Enriched Bins of Su(H) Binding Site Clusters.** Pure bins of Su(H) binding site clusters are the easiest and most obvious to mine for potential target enhancers and the associated genes. However, enriched bins of only moderate purity are also expected to yield valuable candidate targets. To avoid skewing of purity and probability values by contributions from clusters that are also found in pure bins, we removed these regions from the cluster frequency matrix before choosing enriched bins to mine. To this modified data set we applied the same probability and purity analysis procedures described above. In deciding on a definition of enriched bins that would permit efficient mining, we sought to balance frequency and purity, so that a substantial number of candidate clusters could be evaluated, but without undue contamination by randomly expected clusters. We selected a purity cutoff of 50%, and, as before, a probability cutoff of P < 0.05. The specifications of the 36 distinct genomic regions meeting these two criteria are listed in Table 1. As with the pure bins, this list was found to include several binding site clusters (near E(spl)m5, E(spl)m2, and E(spl)m7) that have previously been demonstrated to be functional targets of transcriptional regulation by Su(H) and the Notch pathway (4, 8, 11, 20). Additionally, two other genes that appear on the enriched bin list, numb and neuralized, are closely associated with the function of the Notch pathway. It is highly unlikely that these Notch pathway components
appear on our list by chance, and further investigation of the function of their associated Su(H) binding site clusters is needed.

We were also very interested to note that several clusters on the enriched bin list (Table 1) overlap DNA segments with defined transcriptional enhancer activities that mimic the expression pattern of the adjacent genes. A fragment overlapping three of the four Su(H) binding sites in the cluster that drives expression in the three epidermal attachment cells for larval muscles 21–23 and in the bipolar glia (24). In all three cases, it is plausible that direct regulation by Su(H) could be a component of the enhancer activity.

Functional Analysis of the Su(H) Binding Site Cluster Adjacent to Her. By definition, enriched bins of moderate purity such as we have just considered (Table 1) are expected to be contaminated by false positives, i.e., binding site clusters that appear in the genome by chance and do not represent bona fide enhancer elements. Such contaminated bins may perhaps be mined most effectively through the use of secondary criteria to select candidate clusters for further study. These might include known patterns of expression of nearby genes, if consistent with regulation by the transcription factor of interest, microarray data indicating the possibility of such regulation, or simple gene identity. This last criterion brought to our attention the Su(H) binding site cluster near the gene Hemoglobin (Her) (Table 1; refs. 13 and 25), which like many known targets of direct activation by Su(H) (see above), encodes a bHLH repressor protein. We selected this cluster for experimental analysis to determine whether it identifies a novel target of Su(H) control (Fig. 3).

In situ hybridization to whole embryos and late third-instar imaginal discs, Her transcripts appear to accumulate at low levels in a broad or ubiquitous pattern (data not shown; ref. 25), inconsistent with specific regulation by Notch signaling and Su(H). Further observation that green fluorescent protein reporter constructs containing the Her promoter and upstream region, and either including or not including the four Su(H) sites, exhibit no detectable activity in vivo, suggesting that the Su(H) binding site cluster is not relevant to Her’s pattern of expression (data not shown). These results suggest that the Su(H) site cluster might instead be involved in the regulation of CG12506, a predicted gene that lies ~4.3 kb upstream of Her and is transcribed in the opposite direction (Fig. 3A). We named this gene Him for its proximity to Her. The predicted protein product of Him is itself of interest: the C-terminal four amino acids are WRPW (Fig. 3A), a motif that recruits the compressor Groshes to a wide variety of bHLH and other repressors (20). Unlike Her, the Him protein is not predicted to include a bHLH domain, suggesting that it may instead be a repressor of a different class, or possibly an adaptor protein that functions as an intermediary between Grosho and a DNA-binding repressor.

We found that Him transcripts accumulate in adult muscle precursor cells in both stage 15–17 embryos and imaginal discs of late third-instar larvae (Fig. 3B and C), in a pattern that strongly resembles the expression of a known Notch/Su(H)-responsive gene, E(spl)m6 (17). In particular, Him transcripts appear in a subset of the aedecephalial cell population of the third-instar wing disc (Fig. 3C); these cells give rise to the adult thoracic musculature (27). We first verified that Him transcript accumulation in the wing disc depends on Su(H) function (Fig. 3D). We then observed that a reporter construct that includes the Him promoter and 3.9 kb of upstream sequence, including the Su(H) binding site cluster that appears on our enriched bin list (Table 1), recapitulates the expression pattern of Him in the wing disc (Fig. 3C and E). This construct also includes seven putative binding sites (CACATG; ref. 28) for the mesodermal bHLH activator Twist (Fig. 3A), which is expressed at a largely uniform level in all aedecephalial cells (29). By contrast, a truncated Him promoter construct that includes 2.1 kb of upstream sequence and lacks the entirety of the Su(H) site cluster (but retains six of the seven Twist sites; Fig. 3A) exhibits a greatly reduced level of expression in the region of the aedecephalial cell population that strongly expresses both E(spl)m6 and Him in a Su(H)-dependent manner (Fig. 3F). Our results strongly suggest that the expression of Him is directly regulated by Su(H) via the binding site cluster revealed by our SCORE analysis.

Discussion

We have described here a computational approach to the identification of transcriptional cis-regulatory modules and the associated genes in whole-genome sequence data. This method is based on the now well-established observation that bona fide enhancer elements in bilaterian genomes are usually discrete modules of compact size that frequently contain a discrete set of binding sites for the same transcription factor (1). By identifying statistically nonrandom densities of putative high-affinity binding sites for the Notch-regulated transcription factor Su(H), we were able to recognize in the fully Drosophila genome (13) a substantial number of both known and highly probable cis-regulatory modules that use this factor, along with the genes they serve. Our experimental analysis of one such novel target enhancer, including four Su(H) sites and associated with the gene Him, fully supports the conclusion that it is a site of direct activation by Su(H).

In designing the SCORE approach to binding site cluster analysis, we sought to incorporate several features that facilitate efficient mining of the genome. First, obtaining an unbiased inventory of site clusters over a wide range of sequence intervals (Fig. 1) permits a detailed examination of site density as a global genomic character. Second, the strong statistical foundation afforded by the Monte Carlo simulations of random binding site clustering allows a rigorous evaluation of the significance of all
<table>
<thead>
<tr>
<th>Evidence*</th>
<th>Name of nearest gene/element</th>
<th>Symbol or CG number</th>
<th>Position of closest site</th>
<th>Bin†</th>
</tr>
</thead>
<tbody>
<tr>
<td>a, c, d</td>
<td>Suppressor of Hairless</td>
<td>Su(H)</td>
<td>(+105)</td>
<td>5/300</td>
</tr>
<tr>
<td>a, b, c</td>
<td>Delta</td>
<td>Delta</td>
<td>(-2055)</td>
<td>5/500</td>
</tr>
<tr>
<td>a, c</td>
<td>Ocho</td>
<td>Ocho</td>
<td>(-239)</td>
<td>5/600</td>
</tr>
<tr>
<td>a, b, c</td>
<td>Enhancer of hairy</td>
<td>Enhancer of hairy</td>
<td>Intron 2</td>
<td>8/6,400</td>
</tr>
</tbody>
</table>

| Enriched  |
|-----------|-----------------------------|---------------------|--------------------------|------|
| a, c, d   | Nervy                      | Nervy               | Intron 1                 | 4/1,400 |
| a, b, c   | Hairy                      | Hairy               | Intron 1                 | 4/1,400 |
| a, b, c   | Hairy                      | Hairy               | Intron 1                 | 4/1,400 |
| a, b, d   | Her                        | Her                 | Intron 1                 | 4/1,400 |

| Pure      |
|-----------|-----------------------------|---------------------|--------------------------|------|
| a, c, d   | Suppressor of Hairless     | Su(H)               | (+105)                   | 5/300 |
| a, b, c   | Delta                      | Delta               | (-2055)                  | 5/500 |
| a, c      | Ocho                       | Ocho                | (-239)                   | 5/600 |
| a, b, c   | Enhancer of hairy         | Enhancer of hairy   | Intron 2                 | 8/6,400 |

Su(H) binding site clusters residing in pure and enriched bins of the cluster frequency matrix (Figs. 1 and 2; see text) are listed according to the identity of the nearest gene. Genes and binding site clusters shown in bold are discussed in the text. The location of the binding site closest to the listed gene is indicated.

Letters in this column indicate the nature of experimental evidence supporting the physiological relevance of the Su(H) sites in the identified cluster (see text for details and references). a: Wild-type expression pattern; b: expression in a Su(H) mutant background; c: in vitro DNA-binding assays; d: binding site-dependent enhancer/promoter activity in vivo. A check (✓) indicates that a genomic DNA fragment from a heterologous species demonstrates an activating effect in vivo that recapitulates some aspect of the normal expression pattern of a nearby gene.

This column denotes for each cluster region the first bin in the cluster frequency matrix that identifies the cluster and meets the criteria for pure or enriched bins (see text).

Cluster frequencies observed in the genome (Figs. 1 and 4). Finally, the concept of bin parity (Fig. 2) offers an intuitive, yet quantitative, measure of the degree to which cluster frequencies in the genome exceed the random background, greatly assisting in the choice of favorable bins to mine for potential regulatory targets.

Several parameters contribute to the success of cluster analysis when applied to any given transcription factor, including the quality of the binding site definition, the number of sites in the genome, and whether the factor uses site clustering to any degree to evoke its regulatory response. Su(H) may be particularly favorable in this regard. Nevertheless, we have had considerable success in applying SCORE analysis to Drosophila DNA-binding proteins other than Su(H), including the "properal" 5'HLH activators encoded by the achete-scute complex. Enriched bins (parity ≥50%, P < 0.05) from the properal activator inventory identify binding site clusters located near logical targets of direct regulation by these factors. Genes such as DTRAF1 (30) and wormi (31), both expressed in the
Chapter 2 Supplemental Figure:

34

developing nervous system, are associated with strong concentrations of proneural protein binding sites. We have also found that site cluster analysis can be applied successfully to cis-regulatory elements involved in posttranscriptional regulation, such as the negatively acting 3' untranslated region motifs our laboratory has characterized previously (32-34). These observations suggest that SCORE will be of quite general utility in mining genome sequence data for potential targets of multiple types of regulation.

A clear limitation of the SCORE method when applied to a single transcription factor is that it will generally fail to draw attention to enhancer modules and target genes that use only a single binding site for that factor, or that include more than one site with a statistically random spacing. However, this difficulty can be overcome at least in part by conducting a SCORE analysis with binding sites for more than one factor. The contribution of a single Su(H) site may become significant if this site is part of a statistically unusual cluster of sites for multiple factors. Our knowledge of frequently used combinations of transcription factors, and the expression specificities they control, is growing rapidly (1), making multifactor SCORE increasingly valuable and feasible. For example, our survey of statistically improbable clusters that include binding sites for both proneural HhLH activators (RACAGSTG) and HhLH repressors (CACCGYG) has identified a potential cis-regulatory module in an intron of nvy, which encodes a transcription factor that is the fly homolog of the human oncogenic protein ETF. nvy is expressed in both neuroblasts and sensory organ precursors, a specificity fully consistent with direct regulation by a combination of proneural activators and Notch-regulated HhLH repressors. Significant success in multifactor clustering analysis has also been reported recently by Berman et al. (35).

The data presented in this article demonstrate that the fly genome exhibits widespread and highly significant clustering of binding sites for the transcription factor Su(H) and indicate that cluster analysis can be a sensitive detector of cis-regulatory modules and the associated target genes. As high-quality definitions of transcription factor binding sites and other cis-regulatory sequence elements become increasingly available, SCORE and other similar techniques will no doubt prove increasingly valuable as tools for reading the regulatory genome.

We are grateful to Elizabeth Blankenhorn for suggesting the Monte Carlo simulations. We thank Scott Barolo and Matt Romshaugen for critical comments on the manuscript. This work was supported by National Institutes of Health Grants GM46993 and GM62279 to J.W.P.

Fig. 4. Su(H) binding sites are highly significantly clustered in the fly genome. (A–C) Three-dimensional probability plots of Su(H) site clustering found in the real genome (A), a typical random genome (B), and the most extreme random genome of 50 (C). Number of sites in a cluster is plotted on the x axis; window size (in bp) is on the y axis. Plotted on the z axis, and color-coded according to the key, are probabilities of randomly obtaining a cluster frequency equal to or greater than that observed in the genome analyzed.

Chapter Two, in full, is a reprint of the following publication. Rebeiz M, Reeves NL, Posakony JW. SCORE: a computational approach to the identification of cis-


regulatory modules and target genes in whole-genome sequence data. *Proc Natl Acad Sci U S A.* 2002 Jul 23;99(15):9888-93. I am the primary researcher and author and James W. Posakony directed and supervised the research that forms the basis for this chapter. Characterization of the *Him* gene through GFP reporter and *in situ* hybridization experiments were performed by Nick L. Reeves.
CHAPTER 3
NOTCH REGULATES NUMB: INTEGRATION OF CONDITIONAL AND AUTONOMOUS CELL FATE SPECIFICATION MECHANISMS
Summary

The Notch pathway is a regulatory cassette utilized throughout metazoan development. One conserved component of this cell-cell communication system is the protein Numb, which is characteristically segregated asymmetrically during mitosis, and has a role in antagonizing Notch signal transduction in multiple animal development systems. The purpose of Numb’s role in Notch signaling is best understood in the context of sensory bristle development in Drosophila melanogaster. Numb segregation to one daughter of the primary bristle precursor antagonizes Notch signaling, promoting the default fate in one daughter, and allowing the Notch dependent fate in its sister cell. In a genomic survey for potential Notch targets, the numb gene was identified to contain a statistically significant cluster of Suppressor of Hairless binding sites. Investigating this region in the second intron of numb, we have discovered a Notch responsive element responsible for the production of numb in the pIIa and pIIIb cells of the bristle lineage. In both of these settings, the cell has not received Numb during mitosis, yet must generate sufficient levels to bias cell fates during a subsequent mitosis. Therefore, we propose that while Numb regulation of Notch is an important facet of Notch signaling, Notch regulation of Numb is also a critical component in this elaborated cell lineage setting.
**Introduction**

A fundamental question in developmental biology is how complex transcriptional networks are wired to orchestrate the monumental changes that occur as cells are differentiated from one another. The cellular host of fate determinants and transcription factors co-operate with signaling pathways to modify the transcriptional state of hundreds, if not thousands of genes. As new methods arise to bring these transcriptional networks into a whole genome perspective, we are handed the promise of new biological insights at the click of a button.

The *Drosophila* body is covered with sensory bristles, each of which is a multicellular organ derived by a lineage (Hartenstein and Posakony, 1989). The bristle lineage originates from an epidermal cell that is selected to become a bristle precursor. Once selected, the precursor divides several times, and the progeny of each division adopts a distinct cell fate. The Notch pathway plays a pivotal role both in the selection of bristle precursors, but also in creating cell diversity within the bristle lineage (Hartenstein and Posakony, 1990). A temperature sensitive allele of Notch can disrupt sister cell fates within the bristle lineage such that two sister cells that would normally adopt different fates resultanty adopt identical fates. The protein Numb has been identified as an antagonist of Notch signaling (Guo et al., 1996), and is required to specify sister fates that are opposite of those promoted by Notch.
(Uemura et al., 1989; Wang et al., 1997). The finding that Numb protein is asymmetrically segregated during mitotic divisions of the bristle lineage has provided numerous insights into how two equivalent cells use the Notch pathway to segregate into separate cell fate programs.

A lingering question in the Notch-Numb system has been the origin of Numb protein in a cell that did not inherit Numb from its predecessor (Rhyu et al., 1994). When one confronts this problem, it becomes clear that this is not a straightforward task. The cell that didn’t inherit Numb must be stably specified through the Notch pathway, but is immediately set to the task of setting up a new complement of Numb protein to bias its next division. It has been observed that most epidermal cells express Numb at low levels (Rhyu et al., 1994), however, these levels may not provide enough new Numb protein quickly enough to establish the next round of asymmetric cell fates. Therefore, a careful orchestration of Numb production may be carried out which integrates these conditional and autonomous modes of cell fate specification.

In a previous study, we performed an in silico screen of the Drosophila melanogaster genome for targets of the Notch pathway (Rebeiz et al., 2002). By identifying statistically significant clusters of binding sites for Suppressor of Hairless [Su(H)], we demonstrated that many of the best-known targets exhibit extremely unlikely clusters of binding sites due to random chance. In addition to the known Notch targets identified by the study, several potential targets
were found that already have a highly integrated role in Notch mediated developmental decisions, including *Delta*, *neuralized*, and *numb*. Here we investigate the cluster of Su(H) binding sites in the second intron of *numb*, revealing that *numb* transcription is indeed stimulated in bristle precursor cells in response to Notch signaling. Through this enhancer, an unlikely balance is struck, where the Notch pathway must engage in the task of controlling a potent antagonist of itself. This finding sets up an important precedent for the control of *numb* by Notch signaling.

**Materials and Methods**

**Fly Stocks**

* w^118 is a spontaneous deletion of the *white* locus that eliminates gene function. Sca-Gal4 is an enhancer trap P-element insertion into the scabrous locus, driving GAL4 (Hinz et al., 1994). UAS-Numb flies contain a numb coding region transgene (Wang et al., 1997) that can be misexpressed using the GAL4::UAS system (Brand and Perrimon, 1993).

**Plasmid cloning**

All reporter fragments were amplified by PCR on genomic DNA. Primers used for this study are supplied in supplemental material (Table 3.S1). Site mutants were created by overlap-extension PCR (Ho et al., 1989). The
CD2 fragment and all mutant variants thereof were PCR cloned from the Celera sequencing strain, and are completely compliant with the Celera genome sequence (Adams et al., 2000). Reporter fragments were inserted into *Kpn I* and *Bgl II* restriction sites of the Hsp70 promoter containing GFP reporter pHStinger (Barolo et al., 2000a).

**Germline transformation**

P-element mediated germline transformation was performed as previously described (Rubin and Spradling, 1982), using *w^118* as the recipient strain.

**Antibody staining**

Staining of pupal nota was performed on animals incubated at 25°C, and dissected at timed stages measured in hours after puparium formation (APF). Timed pupae were fixed in PBS + .1% Triton-X-100 (Sigma), 4% Paraformaldehyde (Electron Microscopy Sciences, Fort Washington, PA). For stains using the Hamlet antibody, the Triton concentration in the fixative solution was raised to .3% to increase tissue permeability. Primary antibodies used in this study were mouse-anti-cut (1:100; Developmental Studies Hybridoma Bank; Blochlinger et al., 1990), and guinea pig anti-Hamlet (1:1000; Moore et al., 2002). Secondary antibodies used were goat anti-
mouse Alexa-647, goat anti-mouse Alexa-555, and goat anti-guinea pig Alexa-555 (1:400; Invitrogen). All fluorescent stains were imaged on a Leica confocal microscope. Before both primary and secondary antibody incubations, samples were blocked for 1 hour in a 1:10 solution of western blocking reagent (Roche) in PBS+1% Triton-X-100.

**Mobility shift assays**

Electrophoretic Mobility Shift Assays (EMSAs) using GST-Su(H) were performed as previously described (Bailey and Posakony, 1995). Probe sequences are supplied as supplemental material (Table 3.S1).

**in situ hybridization**

An 1800bp PCR-cloned fragment overlapping the last exon of *numb* was cloned into EcoRI and NotI sites of pBluescript SK I, and used to transcribe a digoxygenin labeled riboprobe (see supplement for PCR primer sequences). In situ hybridization was performed as previously described (Reeves and Posakony, 2005) with the only modification that instead of a normal 1:1500 dilution of 10mg/ml proteinase K, a 1:25000 dilution was used to protect protein epitopes for the detection of the cut protein by antibody stain (above) after the *in situ* was stained.
Sequence searches

The numb orthologous region was retrieved from other Drosophila species using a combination of resources, including the DrosSpeGe Blast server (http://insects.eugenes.org/species/blast/), and the NCBI trace archive (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?). The GenePalette software tool (Rebeiz and Posakony, 2004) was used to manipulate orthologous sequences, and make schematic diagrams. A Perl script named “scrabble” was used to identify the largest conserved blocks across the set of orthologous sequences (script available upon request).

Results

Experimental verification of in silico identified binding sites in numb

In order to preliminarily assess the validity of the putative numb regulatory sequence identified by site clustering over random expectation (SCORE) analysis, we took two initial approaches: electrophoretic mobility shift assay (EMSA) and phylogenetic footprinting (Fig. 3.1B,C). A critical component to any genome survey of transcription factor binding sites is the ability to accurately identify sites that specifically bind the factor of interest. The SCORE survey of the genome used a particularly stringent definition of Su(H) binding consensus, requiring a sequence that matches YGTGDGAA (all IUPAC variants except TGTGTGAA). From previous studies of Su(H) targets,
these sequences are predicted to bind Su(H) protein with high affinity (Bailey and Posakony, 1995; Nellesen et al., 1999; Tun et al., 1994). This prediction is borne out, as all five sites contained in the region bound Su(H) with comparable strength to a positive control probe from the Enhancer of split m4 gene (Fig 3.1B). Furthermore, a single base mutation was sufficient to abolish the majority of binding by GST-Su(H) (Fig 3.1B, “S2mut”). This EMSA data corroborates the accuracy of our Su(H) binding site search criterion, and establishes that the Su(H) matches found in the numb gene are indeed Su(H) binding sites.

Often times, phylogenetic footprinting is used as a measure of biological relevance for a sequence of interest. Although a well-defined cis-regulatory region may exhibit poor conservation of individual binding sites (Ludwig et al., 2000), maintenance of sites over longer evolutionary distances serves as a convenient positive indicator that the sequences are important for regulation. Upon examination of the orthologous region from eight other Drosophila species, three of the five sites are absolutely conserved in all species tested (Fig. 3.1C). A fourth site (S1) is well conserved, being absent in the sister pair D. virilis, and D. mojavensis. Given this site’s conservation across the other Drosophila species tested, it is most likely that this site was recently lost in an ancestor of the virilis-mojavensis lineage. The fifth site (S1b) is extremely close to the well conserved S1 site, and shows conservation only
among other members of the *melanogaster* subgroup. This site lies only eight bp away from the S1 site, a distance that would most likely preclude simultaneous occupancy of both sites. Furthermore, S1 and S1b appear to be related, sharing 11/11 bp of sequence (not shown). This footprinting analysis suggests that four of the five sites have withstood selection for 40-60 million years of evolution, implying an important role for these sequences. Given their Su(H) binding properties, clustering, and conservation, it appears extremely likely that these sequences are physiologically relevant Su(H) binding sites.

**numb expression in developing bristle organs**

The transcriptional regulation of *numb* has not been previously studied, since most attention has been drawn to its protein localization during mitosis (Rhyu et al., 1994). By northern blot, maternal *numb* transcript is present at 0-3 hr after egg laying (AEL), while the zygotic transcript doesn’t become intense until 3-6h AEL, coinciding with the beginning of embryonic SOP divisions (Uemura et al., 1989). By antibody stain, Numb is observed in the SOP and its progeny, but also in surrounding epidermal cells (Rhyu et al., 1994). The inheritance pattern of Numb protein would dictate that some cells in the lineage that didn’t inherit Numb will have to make new Numb before they divide. The only cells that face this predicament are the pIIa and pIIIb cells. Although the pIIb cell must also segregate Numb to its pIIIb sib daughter (Van
De Bor and Giangrande, 2001), the pIIb cell inherits Numb protein from the SOP (Rhyu et al., 1994), and therefore could possibly recycle inherited Numb. Nonetheless, the pIIa and pIIIb cells are faced with task of creating more Numb protein. To determine whether numb transcript accumulates in developing bristles, we performed in situ hybridization on dissected pupae at 16 hours after puparium formation (APF), a time when most microchaete organs are at the 2-3 cell stage of bristle development (Fig. 3.2). At this time, evenly spaced cells or cell clusters specifically express elevated levels of numb. The spacing of these strongly expressing cells is reminiscent of a microchaete field pattern. In agreement with antibody data, numb transcript is also observed at low levels across the epidermal field (Fig. 3.2A).

To assess whether the strongly expressing cells belong to microchaete organs, in situ stained pupae were also labeled fluorescently with the antibody cut, which recognizes nuclei of all developing external sensory organ cells (Blochlinger et al., 1990). Based on the stereotypical divisions of microchaete precursors along the anterior/posterior axis of the notum, we can determine the cellular identity of all microchaete cells during the 1-3 cell stages (Gho et al., 1999). Microchaete positions containing a single cut-positive cell are interpreted to be at the SOP stage of development. Once the SOP divides within the plane of the epithelium, the anterior progeny cell is pIIb, and the posterior is pIIa. Next, the pIIb cell divides perpendicular to the plane of the
epithelium, generating a small, basal pIIb sib cell, and a larger, apical pIIIb cell. The pIIa cell divides next, within the plane of the epithelium, along the A/P axis, and generates socket and shaft cells. In 90% of cases, the pIIa cell divides before the pIIIb sib migrates away, and undergoes apoptosis (Fichelson and Gho, 2003). Therefore, in a typical 2 or 3 cell position it is possible to distinguish pIIa vs. pIIb or pIIa vs. pIIIb vs. pIIIb sib, respectively, based on A/P axis position, size, and apico-basal positioning. A number of cut-labeled SOP positions within the numb-stained notum showed only the background levels of numb transcript accumulation. However, a number of two cell positions stained strongly for numb transcript (Fig. 2B – B’). In all cases examined, the numb positive cell was the posterior-most cell of the organ, thus identified as the pIIa cell. In a subset of three-cell positions, an apical, anterior, normal-sized cut positive nucleus was surrounded by cytoplasmic numb transcript (Fig 2C-C”). Based on the size, location, and position of this cell, we interpret this cell to be the pIIIb cell. Therefore, strong transcript accumulation for numb is observed in the pIIa cell and possibly the pIIIb cell, and not in the SOP or pIIb cells. These expression studies demonstrate that numb has a dynamic pattern of transcript accumulation in the developing bristle organ, which is presumably controlled transcriptionally.

Transcriptional regulatory activity of the numb intron region
As a test of the hypothesis that the Su(H) binding sites found in the *numb* gene are indeed physiologically active, we constructed a 2.5 kb GFP reporter construct (Fig 3.1A) which contains all five Su(H) binding sites identified *in silico*. In independent lines of this construct, we observe expression in the embryonic CNS, PNS, larval CNS, optic lobes, retinal field (not shown) and in bristle precursors of the notum microchaetes (Fig 3.3A).

Based on our phylogenetic analysis of this region, and with the intent to focus efforts on a minimal piece with expression in bristle precursors, the 2.5 kb *numb* region was subdivided into two smaller fragments (CD1 and CD2), each of which contained two Su(H) binding sites (Fig 1A). The 387 bp CD1 fragment maintained expression in the embryonic CNS (Supplemental Fig. 3.S1), but retained no activity in the microchaete field (Fig. 3.3B). However, the CD2 fragment, spanning 682 bp, and containing two completely conserved Su(H) binding sites maintains activity in the microchaete field (Fig. 3.3C), as well as and embryonic CNS (Supplemental Fig. 3.S1), the retinal field, and larval brain (not shown). The retinal field expression of this fragment will be investigated in a separate report (B. Mollereau, personal communication).

Given the broad range of specificities contained in this region, as assayed by our reporter constructs, we conclude that this region of the *numb* gene, identified by an *in silico* search for Notch targets, is a bustling regulatory region of the gene.
Detailed analysis of the CD2 enhancer during bristle morphogenesis

As with the *in situ* analysis of *numb* transcript accumulation, we can also garner exact details on the cellular activation profile of the CD2 enhancer region during bristle development. Using the same criteria for cell fate identification stated above, we performed a detailed analysis of GFP accumulation in 16-18h APF nota of flies carrying the CD2 GFP reporter.

During the SOP stage, no GFP is detected, even when looking at SOPs during their division (Fig. 3.3D). At the two-cell stage, many organs express nuclear GFP in the posterior pIIa cell (Fig 3.3E). Additionally, many two-cell positions have no GFP, consistent with a lag between GFP transcription, and the observation of GFP fluorescence. Because we never see GFP in the anterior cell of a two cell pair, we conclude that GFP transcription is stimulated in pIIa, and not the SOP. Fig. 3.3F shows a later-stage two cell position, in which the pIIb cell is about to divide. In this position, the intensity of GFP in pIIa has increased in comparison to the early two cell position presented in Fig. 3.3E. Notably, even as the pIIb cell is dividing, no detectible GFP is observed in this cell (Fig 3.3F). As the GFP-positive pIIa cell divides, GFP is evenly distributed to its daughter cells (Fig. 3.3G, H). In Fig. 3.3G, a three-cell position is shown, where the posterior pIIa cell is about to divide, as shown by the diffuse stains of both cut and nls-GFP. In Fig. 3.3H, the pIIa cell has divided, and the
presumptive socket and shaft cells have perduring GFP fluorescence from the enhancer’s activity in the pIIa cell. Also, at the four cell stage, an anterior cut-positive cell has also started accumulating GFP (Fig. 3.3H). Based on its anterior location, size, and apical disposition, this cell is putatively the pIIIb cell. The transcription factor Hamlet has recently been shown to be a high-level regulator of the pIIIb cell, and a Hamlet antibody specifically labels pIIIb cells during microchaete development (Moore et al., 2002). Staining 17h APF nota with the Hamlet antibody confirms that this GFP positive cell is indeed pIIIb (Fig. 3.3J). Once the pIIIb cell divides, GFP is inherited by the pIIIb progeny, the presumptive sheath and neuron. At the end of the bristle lineage divisions, a total of four GFP positive cells are observed, caused by the initial expression via the CD2 enhancer in the pIIa and pIIIb cells. Obscured by perdurance of GFP, it is unclear whether postmitotic cells of the bristle organ also the enhancer. However, there appears to be a fairly even expression of GFP between sister cells at later time points (not shown). It is also clear from this analysis that the pIIIb sib cell, and its ancestors (pIIb, SOP) within the lineage never activated the CD2 enhancer, since no GFP is ever observed in this cell. The locations that do activate novel transcription via the enhancer(pIIa and pIIIb) are Notch responsive cells, consistent with the Su(H) binding sites present in the enhancer.
Activation of the CD2 enhancer is Notch dependent

The enhancer analysis of this region matches up nicely to the expression data, and the cis-regulatory logic, dictating that the CD2 enhancer fragment is activated through a canonical Delta/Notch/Su(H) pathway, as suggested by the high affinity Su(H) binding sites found within the enhancer. This model that Notch signaling is required to stimulate numb transcription through the CD2 region has the clear prediction that removal of Notch stimuli from pIIa/pIIb pairs would abrogate enhancer activation. Therefore, in order to test the activity of the CD2 enhancer in the absence of Notch function, we used the GAL4:UAS system (Brand and Perrimon, 1993). Driven by the strongly expressing scabrous-Gal4 driver (Hinz et al., 1994), we overexpressed a UAS-numb transgene in SOP cells. In all documented cases, this treatment shows the specific phenotype of disrupting Notch signaling (Frise et al., 1996; Van De Bor and Giangrande, 2001; Wang et al., 1997). Furthermore, epistasis experiments suggest that the only functions of numb during neurogenesis require Notch function (Guo et al., 1996). Consistent with previous studies, we observe that the adult scabrous-GAL4:UAS-numb flies (Fig. 3.4A-B) are almost completely devoid of external bristle structures on the notum, suggesting a pIIa to pIIb transformation (see Fig. 3.3K for lineage diagram). Staining numb overexpressing flies with anti-cut to label all microchaete cell nuclei at 18h APF shows a relatively normal
pattern of microchaete bristle cells (Fig. 3.4F) comparable to flies carrying only the UAS-*numb* transgene (Fig. 3.4E). However, the CD2 enhancer fragment shows a drastic decrease in GFP expression under these conditions of Notch inactivation (Fig. 3.4D), as compared to UAS-*numb* alone treatment (Fig. 3.4C). These results are consistent with a required input from Notch to the activation of the CD2 enhancer in pIIa and pIIIb cells, both of which show robust expression under wild-type conditions in Fig. 3.4C.

**Dissection of conserved elements within the CD2 enhancer**

To further understand how this novel Notch responsive enhancer region works, we performed a detailed phylogenetic footprinting of the CD2 region, comparing the region from *Drosophila melanogaster* to orthologous regions from eight other species (Fig. 3.5). Both Su(H) binding sites show high levels of conservation at the sequence level (Fig. 3.5, blue boxes). The first Su(H) site (S3, Fig. 3.1C) is conserved completely in a block of 9 bases across eight of the nine species tested. In the ninth species, *D. virilis*, this block of 9 identical bases is disrupted by a single substitution in an allowed wobble position of the Su(H) binding consensus. The second Su(H) site (S4, Fig. 3.1C), is perfectly conserved within a block of 9 of 9 identical bases across all species tested.
The *Drosophila melanogaster* CD2 region contains 4 ETS binding consensuses matching MGGAWRY (Sharrocks et al., 1997). These ETS matches display a range of conservation profiles (Fig. 3.5, gray boxes). The first ETS site (most 3-prime) shows a poor level of conservation, with several differences within the site match among the different species. However, it is striking that in seven of the eight additional species, an occurrence of an ETS match is present in the general vicinity of the first Su(H) site (blue box): It is 59 bp away in *virilis* to 213 bp apart in *ananassae*. Interestingly, the orthology of the site is absolutely unmistaken in certain pairs of organisms (Fig 5, solid lines connecting gray boxes). For example, between *D. ananassae*, and *D. pseudoobscura*, the ETS site in the orthologous position is a seven of seven match, while the spacing between these sites relative to the nearest Su(H) site has changed considerably. The only species lacking an ETS site in this upstream position within the enhancer is *D. grimshawi*, which nevertheless contains 2 well-conserved matches to the consensus in a downstream location. The second ETS site shows poor conservation, and is only present in the very-closely related species to *melanogaster*. The last two ETS sites contained within the *melanogaster* CD2 region display a much higher degree of conservation. In all species searched, at least one of the pair of sites is conserved. This inverted pair of sites shows a spacing of 10 nucleotides in *D. melanogaster*, and this spacing is preserved in six of the eight additional
species searched (Fig 3.5). Interestingly, in *D. erecta*, the downstream site has acquired a mismatch to the ETS consensus, and the site spacing has increased to 12. In *D. ananassae*, the upstream site contains two base changes that are predicted to abolish binding. Finally, in *D. willistoni*, the region between the first ETS site of this pair, and the Su(H) site has been deleted, resulting in an ETS site that overlaps the Su(H) S4 site by one nucleotide. It is interesting to note that *willistoni* contains three other ETS sites, one of which is a strong match by sequence and relative position to the conserved ETS sites found only in *melanogaster, yakuba, and erecta* (Fig 3.5, curved dotted line). All of this evidence points to the possibility that some ETS factor has a role in the regulation of the CD2 region. Furthermore, it suggests that the ETS sites are allowed to come and go more frequently than Su(H) sites, and that perhaps there is stabilizing selection (Romano and Wray, 2003) to maintain some number or loose configuration of ETS sites in this enhancer.

The *melanogaster* CD2 region contains a single match to a relaxed E2F consensus (GCGSSAAA, Fig. 5, red boxes). Although the match found in the CD2 enhancer contains a mismatch to this consensus, the CD2 variant found (GCGGGAAT) has been shown to bind E2F in vitro, and contain functional activity in the polymerase α gene (Yamaguchi, 1997). Six of the additional eight species searched contain an E2F match in the same position. However, in *grimshawi, mojavensis*, and *virilis*, the position at which the mismatch is
found has been shifted. In *D. willistoni*, no match to E2F is found in the orthologous location, and there appear to be no other compensatory E2F occurrences. In *D. ananassae*, an additional mismatch to the E2F consensus is observed. However, there are 3 other matches to the E2F consensus that could compensate for the degeneration of this moderately conserved instance. Based on conservation alone, the importance of an E2F binding site is unclear. It appears to be well conserved, but this conservation may be the artifact of an adjacent conserved region (Mystery Block 2). Also, it is quite possible that most *Drosophila* CD2 regions make use of an E2F binding site, but that the lineage leading up to *willistoni* has discarded this site, and has found a different way to achieve proper regulation.

Two regions show a remarkable degree of conservation, and as of yet do not obviously conform to a known class of binding site. The first region, mystery block 1 (Figure 3.5, green box), is a seven of seven match across all nine species, and maintains a central position between the two absolutely conserved Su(H) sites. In seven of nine cases, it forms a perfect match of 9 nucleotides. The second mystery block region, (Fig. 3.5, “Mystery Block 2”), shows an extended region of loose homology. Over a span of 22 nucleotides, 16 nucleotides match perfectly in groupings of 3 to 5 bases. One conspicuous aspect of this region is a perfectly conserved 5 bp stretch that contains a canonical hox consensus ATTA (Pearson et al., 2005). Given their absolute
conservation, the function and binding partners for these mystery sites may provide important insights into the operation of this enhancer.

**Direct binding of Su(H) is required for normal CD2 enhancer activity**

To address the question of which sites are important for proper regulation of the CD2 enhancer, we made site mutant versions of the CD2 enhancer. A two-nucleotide mutation each of the pair of most conserved ETS binding sites had little to no impact on enhancer activation in pIIa (Fig. 3.6B-B”). Likewise, normal pIIa expression was observed when the E2F and Mystery block 1 sites were mutated (Fig 3.6C”-D-D”). However, when a single nucleotide modification was made to the Su(H) binding sites (YGTGDGAA-> YGTGDCAA), expression from the CD2 enhancer is de-repressed, showing expression in both pIIa and pIIb (Fig. 3.6E - E”). This is completely consistent with a role for Su(H) as a repressor in the pIIb cell (Barolo and Posakony, 2002).

**Discussion**

We have described a new and important regulatory linkage between Notch, a key extrinsic cell fate specification pathway, and Numb, an intrinsic cell fate specification protein. A discussion of the implications from this work follows.
Computational searches for enhancers bear fruit

The motivation to study the transcriptional regulation of *numb* originated from a computational search of the fly genome for new Notch targets (Rebeiz et al., 2002). Although a multitude of Notch pathway targets have been documented during SOP selection (Bailey and Posakony, 1995; Lecourtois and Schweisguth, 1998; Nellesen, 1998), the list of bristle lineage Notch targets is quite scant (Barolo et al., 2000b; Nellesen et al., 1999). Therefore, any method to enrich potential targets from the pool of ~14,000 genes in the genome can be a useful route to gaining a greater understanding of Notch function during sister-cell fate decisions. An impressive side-effect of this type of search is the finding that these approaches can lead the operator to think about a regulatory linkage that they normally would have dismissed immediately – the regulation of *numb* by Su(H) seems almost preposterous at first.

One problem associated with finding enhancer sequences is that they can act at great distances to modulate the transcriptional activity of a locus. Within the *numb* locus, there is ~52kb of potential regulatory sequences, making the traditional approach of overlapping 5kb fragments an impossible chore. As clearly demonstrated by this enhancer, the strategy of looking directly upstream of the transcription start site would also have led one astray
from the path to this regulatory element. One beneficial feature of the SCORE approach was the use of a fairly unbiased cluster size to find statistically significant Su(H) binding site clusters. Many studies have used cluster sizes of 500-1000bp, because this adheres to the canonical view of an enhancer element (Davidson, 2001). When searching, we looked through a series of cluster lengths, ranging from 100bp to 5000bp. This wide range may allow one to detect a local maximum that doesn’t necessarily match the size of a canonical cis-regulatory module. Judging from this study, and many of the other clusters found in our analysis, the unbiased cluster-size approach may lead one to discover functional enhancer elements due to the fact that multiple enhancers with similar binding inputs are themselves clustered together.

The role of Notch in de-repressing targets during lineage decisions

Our enhancer studies suggest that Su(H) plays only a minor role (if any) in the activation of the CD2 enhancer (Fig 3.7). Therefore, we are left with the model that other potent activators are present in both pIIa and pIIb, and that Su(H) mediated repression is crucial to prevent the precocious activation of numb before the pIIa cell has been specified by Notch signaling (Fig 3.7). This model dictates that the activators present in both cells would be established early in the life of pIIa and pIIb cells. Interestingly, the Su(H) mutant enhancer is not active in SOP cells. Therefore, if these activators are present during
SOP stages, some mechanism must silence the enhancer until after SOP division. Mutations in individual binding sites (MB1, E2F, the conserved ETS site pair) show a normal pattern of GFP accumulation in the pIIa cell. However, if they have a quantitative effect on enhancer output, this effect might not be detectible within the sensitivity of these GFP reporter assays. One confounding problem with p-element mediated transformation is that the insertion site can influence the relative strength of expression from identical constructs. This “line-to-line” variation can hide subtle differences between constructs. Assessment of expression from double mutants might allow the function of these individual sites to be uncovered. Identification of the local activators in this enhancer will allow us to better understand how these transient cells are specified.

**Is Numb recycled during bristle lineage progression?**

The finding of transcriptional regulatory sequences of *numb* that drive expression in the pIIa cell naturally raises the question of whether pIIb must implement an opposing transcriptional loop. However, several scraps of evidence suggest that Numb protein can be recycled during progression of the bristle lineage: (1) The pIIb cell has an extremely short period of time between birth and division. (2) During this period of time, strong transcript accumulation is observed only in the pIIa cell (Fig 3.2B-B’’). One would predict that if novel
transcription of *numb* was required in pIIb, a comparable level of transcript accumulation would have been observed. (3) Mosaic analysis shows that mitotic clones of *numb* induced during the SOP stage have defects in sheath-neuron decisions as a result of lack of *numb* in the pIIIb cell. However, these clones would not be able to make Numb protein in pIIb, and therefore are predicted to consist entirely of pIIIb sib cells unless a mechanism of Numb recycling is sufficient to promote the pIIIb sib fate. However, an explicit test of Numb protein recycling is technically difficult, and we cannot rule out the possibility that the pIIIb/pIIIb sib decision is biased in other ways such that loss of numb is irrelevant.

The idea of Numb recycling raises an interesting possibility for elaboration of lineages during evolution. If a mechanism for Numb recycling as well as Notch stimulated transcription exists, such as in the CD2 enhancer, one can imagine a mechanism to ensure differentiation of sister cell fates when a new division arises in the lineage: the cell which inherited Numb can now re-use this protein, and the other cell has a mechanism to rapidly transcribe it.

**Notch regulates numb: turning the tables on an old paradigm**

The classical view of Notch and Numb in asymmetric specification of cell fates has focused on the role of Numb in antagonizing Notch signaling to
bias cell fate decisions which would normally turn out symmetrically (Lai, 2004). Our study shows that a reciprocal regulatory loop is also in place: Notch regulates *numb*. The problem of *numb* acquisition within cells that do not inherit *numb* was contemplated by Rhyu in 1994: “the asymmetric division of the IIa cell in the adult external sensory organ lineage also appears to rely on *numb* function…this implies that numb is expressed in both IIa and IIb before division and distributed asymmetrically to their daughter cells” (Rhyu et al., 1994). Although *numb* is generally expressed at low levels in epidermal cells, the rapid divisions of the bristle lineage may not provide enough time for the accumulation of *numb* through a general epidermal enhancer. Activating transcription of *numb* in a Notch responsive cell is a dangerous task, so it makes sense that the Notch pathway would play a major role in presiding over this delicate decision, protecting the pIIa cell fate before allowing *numb* production.

An interesting finding of this enhancer was that it was active in pIIa and pIIIb. It is unclear at the moment if these two cells activate the enhancer through an identical mechanism, or if there are individual controls for each cell. Further mutagenesis of individual sites within the enhancer may shed light on this question if mutations with differential effects are found. This enhancer does represent the first regulatory sequence documented for these cells of the bristle lineage, and adds to the scant list of Notch responsive genes during
sister cell fate specification. The only other Notch responsive enhancers during bristle development we are currently aware of are the Su(H) autoregulatory enhancer in the socket cell (Barolo et al., 2000b), and the $E(spl)\, m\gamma$ enhancer in an undetermined cell of the bristle lineage (Nellesen et al., 1999). The discovery of this regulatory loop in *numb* highlights the possibility that other Notch responsive bristle lineage enhancers may be found on the SCORE list.
Figure 3.1. Validation of putative Su(H) binding sites in numb found by whole genome in silico analysis. (A) Overview of numb genomic region and constructs used in this study. Image was drawn to scale with the position of Su(H) sites marked with an “S”. The boxed area marks the cluster of five binding sites found by SCORE analysis. (B) EMSA assay of sites in Boxed region of (A). Site names are numbered according to naming convention listed in (C). All sites bind a GST-Su(H) protein with comparable strength to the positive control site (m4 s3). A single substitution in S2 nearly completely abolishes binding (S2mut). (C) Conservation of binding site cluster among 9 Drosophila species. Solid lines indicate site orthology based on both the presence of a site in an orthologous location as well as shared sequences near site pairs (See Figure 4 for an alignment of sites S3 and S4). The dotted line represents a pair of site occurrences that are a 10/10 match between D. pseudoobscura and D. mojavensis, but for which matches are not present in other species.
Figure 3.2. *numb* is specifically expressed in the bristle lineage. (A) At 16h APF, *numb* is expressed at discrete positions within the microchaete field, against a background of low-level ubiquitous expression. (B – B’’) *numb* transcript accumulates specifically in the pIIa cell, identified by cut staining (B) of a two-cell stage developing bristle. pIIb is the anterior cell, and pIIa is the posterior cell (see Fig 3K for lineage diagram). (C–C’) *numb* transcript accumulation in a pIIIb cell. (C) cut staining of a three cell stage bristle, where the anterior pIIb cell has divided into the pIIIb and pIIIb sib cells. The larger pIIIb cell accumulates numb transcript, while the smaller, apical pIIIb sib cell does not.
Figure 3.3. Restriction of Su(H) responsive bristle lineage enhancer to a 683 bp fragment active in pIIa and pIIlb cells. (A-C) View of microchaete field at 24h APF, after all bristle cells have divided. (A) The 2.5 kb reporter expresses GFP in multiple cells of the bristle lineage at this time. (B) No GFP is detected in bristle cells of the CD1 387 bp reporter. (C) The CD2 683 bp fragment drives GFP expression identical to the 2.5 kb reporter. (D-K) Stepwise analysis of CD2 reporter during bristle development. Cut (blue) labels all cells of the microchaete lineage. Schematic diagrams show the inferred identity of each cell. At the one cell stage (D), no GFP is present. Once the SOP divides, GFP is present in the posterior pIIa cell (E), and completely absent in the pIIb cell, even as it nears division (F). As the pIIa cell divides (G), GFP is distributed to its daughters, the presumptive socket and shaft cell pair (H). GFP then accumulates in one of the anterior cells, presumably the pIIlb, based on its larger size relative to pIIlb sib (H). Finally, once this anterior cell divides, GFP is distributed to its progeny, the presumptive neuron and sheath cells (I). (J) Hamlet staining of four cell microchaete position, demonstrating that the anterior GFP positive cell is the pIIlb cell, marked by Hamlet. (K) Lineage diagram of notum microchaetes.
Figure 3.4. Notch signaling is required for activity of the CD2 reporter fragment. (A-B) Adult nota of UAS-Numb (A) and Sca-Gal4:UAS-Numb (B) flies. (C-H) 18h APF notum of UAS-Numb (C, E, G) or Sca-Gal4:UAS-Numb (D, F, H) fly. (C) GFP is strongly expressed in all microchaete organs older than the two cell stage in a fly bearing UAS-Numb alone, consistent with the pattern shown in Fig. 3. However, this expression is lost once UAS-Numb is overexpressed in bristle precursors by the Scabrous-Gal4 driver (D). (E,F) Cut staining, and merged representations (G,H) demonstrate that bristle organs are of a correct stage to observe GFP expression under wild-type conditions.
Figure 3.5. Alignment of the CD2 regulatory region across nine *Drosophila* species. In alignments, * indicates residue conserved in all species. Bold characters represent regions of the alignment that correlate with likely regulatory factors (e.g. ETS or E2F), or with regions of high conservation, such as the stringently conserved Mystery block regions. The graphical alignment depicts spatial relationships of the nucleotide alignments. Site matches that occur on the top strand of DNA appear above the DNA line. Reverse complement matches appear beneath the line. Solid lines represent site instances that are clearly related by sequence homology. Dotted lines indicate site instances that are in a similar location, but for which an orthology relationship is difficult to assess. Red circles highlight instances in which a site is not present even though other closely related species contain one in a similar location.
Figure 3.6. Expression of mutant CD2 GFP reporters. (A-E) Two-cell microchaete positions marked by cut (red), compared to GFP expression by CD2 enhancer region variants (A'-E'). (A''-E'') Merged images demonstrate coincidence of GFP and developing bristle nuclei.
Figure 3.7. Model for numb expression in the pIIa cell. (A) In the SOP, Su(H) is a repressor, keeping the numb enhancer off. (B) In newly divided pIIa/pIIb pairs, one or more transcriptional activators (ACT) are counterbalanced by Su(H) mediated repression, keeping numb in the repressed state (OFF). (C) Once a sufficient Notch signal has been received in pIIa, repression by Su(H) is relieved, and the activators can then turn numb transcription ON. This activation of numb may involve some positive input from Su(H). (D) When the Su(H) binding sites are mutated in the enhancer, these putative activators, present in both pIIa and pIIb are weakly sufficient to activate transcription. Green outline represents the inheritance of Numb protein.
Table 3.1 Primers used in the study. For *in situ* probes and reporters, lowercase letters refer to the restriction sites used to clone PCR products into the desired vector (see methods).

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>m4</em> S3</td>
<td>GGCAGCCGGCACGCACAGGG</td>
</tr>
<tr>
<td>numb S1</td>
<td>TCGCGCATGCGCGTGCTTAATGGT</td>
</tr>
<tr>
<td>numb S2</td>
<td>TGGATAGGCGCGTGGCCTTTGGG</td>
</tr>
<tr>
<td>numb S3</td>
<td>TGGATAGGAGAGGGGCTCCTTTGC</td>
</tr>
<tr>
<td>numb S4</td>
<td>CAAAAACGCAGGTGTTTTTTC</td>
</tr>
<tr>
<td>numb S1b</td>
<td>TTTCCTGGCGAGCTGTCGCGCAT</td>
</tr>
</tbody>
</table>

**Emsa Probes**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>numb 2.5-F</td>
<td>ggggtaccccGGCTGCATCTGGTCAAACGACAG</td>
</tr>
<tr>
<td>numb 2.5-R</td>
<td>gaagatcttcTGGAGTAGACACCTCGGTAATCC</td>
</tr>
<tr>
<td>numb CD1-F</td>
<td>gaagatcttcTCTCTAATTTTGATCTCGGACGC</td>
</tr>
<tr>
<td>numb CD1-R</td>
<td>ggggtaccccATCTCTGACATCGCTGCTAGCT</td>
</tr>
<tr>
<td>numb CD2-F</td>
<td>gaagatcttcGATAGTGGATAGTGATGGTGGTTC</td>
</tr>
<tr>
<td>numb CD2-R</td>
<td>ggggtaccccTACCTCAGCTTGAAGGTGTTTTCC</td>
</tr>
</tbody>
</table>

**in situ Probe**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>numb-IS-F</td>
<td>gcgaattcGGATTACTTCATGTCAGCGGACAG</td>
</tr>
<tr>
<td>numb-IS-R</td>
<td>ggccggcgccccGTGGGTAACCTGGTACGTGCA</td>
</tr>
</tbody>
</table>

**Mutagenesis Primers**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>numb Su(H) s3m-F</td>
<td>ACGAAGAGTTGCCACGACCACCAGA</td>
</tr>
<tr>
<td>numb Su(H) s3m-R</td>
<td>TCGGTCGTCGAGACTCTTTGC</td>
</tr>
<tr>
<td>numb Su(H) s4m-F</td>
<td>CACTTCGCGCTGTCAACTTTTCA</td>
</tr>
<tr>
<td>numb Su(H) s4m-R</td>
<td>TTGAAATTTGACAGCGCGGAAGTG</td>
</tr>
<tr>
<td>numb ETS-1mut-F</td>
<td>ATACGCAATTAGTCTCAGCAA</td>
</tr>
<tr>
<td>numb ETS-1mut-R</td>
<td>TTTGTGAGCTTAATTGCGTAT</td>
</tr>
<tr>
<td>numb ETS-2mut-F</td>
<td>GAAGAGATTACTTGGAGGAGA</td>
</tr>
<tr>
<td>numb ETS-2mut-R</td>
<td>TCACTCCGCTAATGTCGCT</td>
</tr>
<tr>
<td>numb ETS-3/4mut-F</td>
<td>AAAAAAAAGAAGAAattAAGCCATTAACATGTTaTTTGTACACTCA</td>
</tr>
<tr>
<td>numb ETS-3/4mut-R</td>
<td>TGTGTCGTTAAAattAATCGTTTTATTAGCTTaTTTCCTTTT</td>
</tr>
<tr>
<td>numb E2F mut-F</td>
<td>AGCGGTATGGCTGGCTGAATTATAA</td>
</tr>
<tr>
<td>numb E2F mut-R</td>
<td>TTATATTACGCAAGCGCATAGC</td>
</tr>
<tr>
<td>numb MysteryB1-F</td>
<td>GAACCTCGATTTTTTTTGAAGAC</td>
</tr>
<tr>
<td>numb MysteryB1-R</td>
<td>GTCTATCATAAAAAATCGAGTTC</td>
</tr>
<tr>
<td>numb MysteryB2-F</td>
<td>CGGAATAAAAAAATTTAAAAACAGTAAGTACAGAAAAA</td>
</tr>
<tr>
<td>numb MysteryB2-R</td>
<td>TTTCCTGAGTGACTTGAATTTTTT</td>
</tr>
</tbody>
</table>
Figure 3.S1. Expression of CD1 and CD2 fragments in embryos. (A, B) The CD1 fragment is expressed in the embryonic CNS. (C, D) The CD2 fragment shows expression in both CNS and PNS. (A) and (C) are lateral views of a stage 9-11 embryo. (B) is a ventral view. (D) is a ventrolateral view of a stage 12 embryo, showing CD2 driven GFP expression in multiple PNS cells as well as in the ventral nerve cord. Both enhancers drive complex expression in the embryonic head.
References


CHAPTER 4
A CODE FOR NEURAL PRECURSORS
Summary

During neurogenesis, proneural genes are required to select single neural precursors from an equivalence group termed the proneural cluster. For years, Notch signaling has been known to have a role in the selection of a single precursor. Mutations in the Notch pathway result in the selection of multiple precursors from the proneural cluster. One main mechanism of Notch pathway action is through the transcriptional upregulation of basic-Helix-Loop-Helix repressors (bHLH-repressors) of the Enhancer of split complex (E(spl)-C). Although this is a well-supported finding, the downstream targets of these repressive proteins are as of yet unknown. In this report, we describe an enhancer in the gene phyllopod, and place it directly downstream of the proneurals achaete and scute, as well as the E(spl)-C repressors. A reporter gene shows de-repression throughout the proneural cluster when Enhancer of split binding sites are mutated. Using this enhancer as a template, we discover bioinformatically one other enhancer that uses a similar logic, within an intron of nervy.

Introduction

Notch is one of the handful of cell-cell communication pathways used throughout metazoa. Not only is the core cassette of Delta (Ligand), Notch (Receptor), and Suppressor of Hairless (Transcription factor) conserved in all
creatures who use the pathway, but the terminal Enhancer of split \([E(spl)]\) basic-Helix-Loop-Helix (bHLH) repressor targets of Suppressor of Hairless are also conserved as part of the Notch functional unit (Lai, 2004). One of the key uses of the Notch pathway is to select a single cell to adopt a distinct cell fate from a pool of equipotent cells. This process of “lateral inhibition” has been observed in several settings, and follows the general paradigm of activating bHLH-repressor genes to inhibit the equipotent pool (Artavanis-Tsakonas et al., 1999).

The *Drosophila* bristle serves as an excellent model to study lateral inhibitory signaling. An important problem during the specification of bristle cell fates is to achieve the proper layout and spacing of these organs to provide the animal with the most effective sensory apparatus possible. During the earliest stages of bristle development, precursors are selected from clusters of equipotent epithelial cells termed the “proneural cluster”, defined by the expression of proneural bHLH-activators *achaete* and *scute* (Cubas et al., 1991; Skeath and Carroll, 1991). Normally, a single sensory organ precursor (SOP) is selected from this field of 20-30 cells. Mutations in the Notch pathway cause the selection of too many SOPs (Hartenstein and Posakony, 1990; Parks and Muskavitch, 1993; Schweisguth et al., 1994; Schweisguth and Posakony, 1994). These SOPs divide, differentiate, and together form a tuft of bristles. Several important targets of the Notch pathway have been identified,
and their regulatory regions have been examined (Bailey and Posakony, 1995; Lecourtois and Schweisguth, 1995; Nellesen et al., 1999). One theme of these lateral inhibitory target modules is that they all follow the similar logic of binding both bHLH activators of the achaete-scute family, but also input from Suppressor of Hairless [Su(H)], the terminal transcription factor of the Notch pathway (Castro et al., 2005). Thus, these enhancers all implement a Su(H) + proneural “code”.

Although a wealth of knowledge is known about these direct targets of the Notch pathway in this setting, extremely little is known about the indirect targets: what genes are directly shut off by the Enhancer of split bHLH-repressors? The only targets implicated thus far are the proneural factors themselves. First, the expression pattern of achaete and scute starts out ubiquitously throughout the proneural cluster, and then refines down to a single cell, the SOP (Cubas et al., 1991; Skeath and Carroll, 1991). Therefore, there must be a mechanism for restricting proneural expression in the cluster.

A high affinity bHLH-R binding site directly upstream of achaete has been implicated in the direct repression of achaete in non bristle forming regions via the bHLH-repressor hairy (Van Doren et al., 1994). This binding site has also been shown to be conserved over extremely long evolutionary distances, and therefore, must provide an important abstract function to the regulation of achaete (Rebeiz et al., 2005). In addition to its role in hairy-mediated
repression, this binding site has also been implicated in the repression of proneurals during SOP selection (Jiménez and Ish-Horowicz, 1997). By misexpression of a VP-16 fusion of \( E(spl) m7 \), it was shown that ectopic bristles can be induced, presumably through the activation of \( E(spl)m7 \) targets. Taking into account that these bristles form in a manner that depends heavily upon the \textit{achaete-scute} gene complex, it suggests that the major contribution to this phenotype is made through the upregulation of proneurals by this fusion protein. However this study did not establish that this phenotype depends on the bHLH-repressor site upstream of \textit{achaete} (Jiménez and Ish-Horowicz, 1997).

With the intent of finding other targets of the lateral inhibitory \( E(spl) \) bHLH-R’s, we have constructed a hypothetical code for SOP genes that complements the proneural + Su(H) code for Notch responsive genes (Fig. 4.1). This code proposes the logic that an SOP gene might have modules that integrate proneural activators and bHLH-repressors concurrently. Through proneural binding sites, the module would have the potential to be activated in any cell of the proneural cluster. This potential is squelched, however, by action of bHLH-repressors, which would turn this enhancer off in any cell that is not the SOP. In the absence of Su(H) inputs or bHLH-repressor inputs, an enhancer may drive ubiquitous expression throughout proneural clusters (Reeves and Posakony, 2005), therefore providing a logical framework for
three different specificities within a proneural cluster (SOP, Non-SOP, and whole cluster).

Using this hypothetical proneural + bHLH-repressor code we have searched through many known SOP genes for combinations of binding sites for these two factors. Within the intron of phyllopod (phyl), a gene required for the SOP fate (Pi et al., 2001), we have found a cluster of binding sites for both of these factors. Using GFP reporter assays, we show that this cluster recapitulates phyl’s native expression pattern in imaginal wing discs. By applying a whole-genome transcription factor binding site strategy (Rebeiz et al., 2002), we show that this cluster exhibits a composition of binding sites that is improbable by random chance. Looking at another cluster found in the same significance class as phyl, we identify an enhancer region of nervy, another SOP gene. This cluster drives expression in the triple row bristle SOPs of the wing margin. Both of these enhancers exhibit de-repression when bHLH-repressor sites are mutant, consistent with the model of our code for neural precursors (Fig. 1). These findings introduce the first genes to be placed downstream of lateral inhibitory bHLH-repressors, as measured through a reporter assay.

Materials and Methods

Fly Stocks
$w^{1118}$ is a spontaneous deletion of the *white* locus that eliminates gene function. $sc^{10-1}$ is a compound mutation that inactivates both *achaete* and *scute* (Flybase consortium, 2003).

**Plasmid Cloning**

All reporter fragments were amplified by PCR on genomic DNA. Primers used for this study are supplied in supplemental material (Table 4.S1). Fragments were PCR cloned from the Celera sequencing strain, and tested for compliance with the Celera genome sequence (Adams et al., 2000). The *phyllopod* intron reporter has one difference in regards to the Celera sequence: a run of 14 A’s is truncated to 13 A’s. Site mutants were created by overlap-extension PCR (Ho et al., 1989) and sequence verified. Reporter fragments were inserted into Kpn I and Bgl II of the Hsp70 promoter containing GFP reporter pHStinger (Barolo et al., 2000).

**Germline transformation**

P-element mediated germline transformation was performed as previously described (Rubin and Spradling, 1982), using $w^{1118}$ as the recipient strain.

**Electrophoretic Mobility Shift Assays**
DNA binding assays were performed as previously described for Achaete/Daughterless heterodimers (Singson, 1995), and m7-truncated (Van Doren et al., 1994). Probe sequences are in Supplemental table 4.S1.

**Sequence searches and informatics**

Using the GenePalette software tool (Rebeiz and Posakony, 2004), we searched known SOP genes for clusters of high affinity proneural and bHLH-repressor binding site consensuses. The proneural binding consensus used was RCAGSTG, defined by a wealth of *in vitro* binding assays (Cabrera and Alonso, 1991; Singson et al., 1994; Van Doren et al., 1991). To find bHLH-repressor sites, we allowed a single mismatch to a high affinity bHLH-repressor binding consensus that is a composite of *E(spl)* and *hairy* binding specificities (Jennings et al., 1999; Rebeiz et al., 2005; Van Doren et al., 1994). Mismatches were only permitted outside of the CACGYG core hexamer, since these nucleotides are critical to *in vitro* binding of *E(spl)* class bHLH-R’s.

The site clustering over random expectation (SCORE) approach was applied to proneural and bHLH-repressor sites across the whole *Drosophila* genome. SCORE analysis was performed as previously described (Rebeiz et al., 2002). A new script was used to calculate co-occurrence clusters of two or more different binding sites (scripts available upon request). Simulations
using the same number of proneural and bHLH-R sites found in the *Drosophila* genome (Release 3) were performed 1000 times. A significance cutoff of $p < .005$ was employed across simulated data sets. See supplemental Table 4.S2 for a list of significant clusters contained in the cluster-class as *phyl*.

**Results**

**Validation of a binding site cluster in *phyllopod***

In a search of known SOP genes for clusters of proneural and bHLH-repressor binding sites, we came across the intron of *phyllopod (phyl)* (Fig. 4.2A). This cluster contains three proneural, and two bHLH-R sites in ~300 bp, and is the most dense collection of both binding sites in the 18 kilobase (kb) territory of the *phyl* gene. Using the phylogenetic footprinting strategy, we observe that these sites are conserved relatively well (Fig. 4.2B, C). Between nine drosophila species spanning at least 40-60 million years, we observe that all three proneural binding sites are maintained in an orthologous spacing and orientation. The repressor binding sites in this region are also well conserved. However, in *Drosophila virilis*, one of these repressor sites is not present.

Overall, the spacing between sites has remained relatively constant. The size of the cluster ranges from 208 bp in *D. ananassae* to 303 bp in *D. grimshawi*. Though sites show a relative maintenance of spacing, none show an absolute constraint on spacing relative to one another.
The nucleotide sequences surrounding the binding sites also exhibit a considerable degree of conservation (Fig. 4.2C). The proneural sites, P1, P2, and P3 (see Fig. 4.2B for numbering) are conserved in blocks of 9/9, 7/7, and 11/11 nucleotides across all nine species, respectively. The bHLH-repressor sites show a more relaxed degree of sequence constraint, as none of them are conserved through the whole consensus used. The R1 site has conserved 7/8 bases within the 10 bp consensus used. The R1 site can be recognized in *D. virilis*, but its novel mutation in this species violates the core hexamer (CACGYG), and is therefore predicted to not bind bHLH-repressor proteins.

The R2 site shows a greater degree of conservation, maintaining a block of 8/8 nucleotides across all species. In 7 of 9 species analyzed, the complete 10 bp sequence is conserved. Interestingly, the repressor sites in this region follow the core hexamer selected for *hairy* (Van Doren et al., 1994) rather than that selected by *E(spl)* bHLH-repressors (Jennings et al., 1999).

The next step in our validation of this cluster was to use Electrophoretic mobility shift assay (EMSA) to assess the *in vitro* binding of these consensus matches (Fig. 4.3). Using heterodimers of *achaete* and *daughterless*, we find that all three proneural binding sites are bound at a comparable strength to a positive control (*m4 E1*, Fig. 4.3A). This binding can be efficiently broken by a two-base mutation to the E-Box (*RCAGSTG-* > *RAAGSGG*, *phyl* P1 mut, Fig. 4.3A). Similarly, the bHLH-R sites within this region can be strongly bound by
a truncated GST-fusion of \( E(spl)m7 \) (Fig. 4.3B), in agreement with our prediction of these binding sites. The bHLH-repressor binding site is sensitive to a thee base mutation that disrupts its E-Box (\( CACGYG \rightarrow CCCTYT \)). Based on these phylogenetic footprinting, and EMSA studies, we conclude that these sequences are most likely functional regulatory sequences due to their stringent pattern of conservation within the region, and that these are proneural and bHLH-R binding sites, given their \textit{in vitro} properties.

**In vivo test of the phyllopod binding site cluster**

To test our prediction that this cluster of binding sites drives proneural cluster expression, we inserted a 662bp fragment (Fig. 4.2A) of the \textit{phyl} intron containing the site cluster into a reporter construct driving GFP (Barolo et al., 2000). This construct drives robust GFP expression within wing proneural territories (Fig. 4.4A'). The reporter nicely recapitulates the native \textit{phyllopod} expression pattern, showing strong expression in SOPs and weaker expression in non-SOP cells (Reeves and Posakony, 2005). This finding is consistent with our model of a dual input code for neural precursors (Fig. 4.1).

**Expression of mutant reporters**

In order to explicitly test our dual input model (Fig. 4.1), we made mutations of the proneural and bHLH-R binding sites within the 662 bp
enhancer fragment (Fig. 4.4). To facilitate the analysis of mutant reporters, we used the GFP-RFP reporter system (Barolo et al., 2004) to compare wildtype and mutant reporters in the same tissue simultaneously. First, to validate the system, we tested wildtype *phyl* reporters driving both GFP and RFP at the same time (Fig. 4.4A-A”). In this case, the expression pattern of GFP and RFP is nearly indistinguishable, other than a slight increase in GFP strength relative to RFP. A comparison of four individual wildtype GFP lines to the wildtype RFP line shows that all wildtype lines behave similarly (Fig. 4.S1), validating the GFP:RFP comparison for this enhancer.

Because bHLH-repressors are predicted to restrict expression from non-SOP cells, a reporter with mutant bHLH-repressor sites is predicted to show expansion into the non-SOP cells of the proneural cluster. When the bHLH-R sites are mutated in the context of the 662 bp GFP reporter (Fig. 4.4B-B”), an extraordinary expansion of GFP relative to RFP is observed, consistent with the prediction of the dual input model (Fig. 4.1). Robust expansion relative to the wildtype reporter is observed in all wing disc proneural clusters though it is most pronounced in scutellar and dorsocentral clusters, and less obvious in the wing margin, and notopleural clusters. This behavior of the bHLH-repressor mutant reporter is consistent across four independent insertion lines tested (Fig 4.S2).
The proneural sites are predicted to provide a positive input to the enhancer (Fig. 4.1), therefore mutation of these sites should decrease activation through the enhancer. This prediction is fulfilled when comparing the wildtype RFP reporter to the proneural mutant GFP reporter (Fig. 4.4C'). Although GFP expression is still present in SOPs, levels of GFP are decreased relative to the wildtype reporter (compare Fig. 4.4C' to 4.4A'). Additionally, there is extra expression of GFP throughout the wing pouch, which is reproduced in each of the 4 GFP lines tested (Fig. 4.S3). Interestingly, some positions are completely absent, including the *atonic* dependent dorsal radius SOPs.

Both types of mutations to the 662 bp fragment are consistent with our proposed code for SOP modules. The predicted code not only was successful in helping us find a new SOP module, but also proved to be predictive of function when sites were mutated in the context of this module.

**The *phy1* module is downstream of proneural genes**

Because the proneural mutant reporter maintains some activity in SOP cells, it seems likely that an additional factor contributes to activation in these cells. This contrasts slightly with the finding that *phy1* transcript is absent in a *sc*^{10-1} background which is devoid of *achaete* and *scute*. To resolve this issue, we assessed the activity of the wildtype 662 bp GFP reporter in a *sc*^{10-1}
background (Fig 4.S4). Under these conditions the imaginal wing disc is
devoid of GFP, except in the position of the dorsal radius, which produces
sensilla that depend on the proneural bHLH activator *atonal* (Fig. 4.S4A).
*atonal* dependent GFP expression in the eye-antennal disc was unaffected, as
predicted (Fig. 4.S4B). Therefore, the residual expression in the proneural
mutant *phyl* reporter must be indirectly downstream of proneural function.

**The *phyl* module is statistically significant in whole genome analysis**

The cluster of proneural and bHLH-repressor binding sites within the
*phyl* genomic region stands out when scanning its 18kb genomic locus (Fig
4.2A). Because of its anomalous appearance, as well as the apparent
effectiveness of using this code to find the *phyl* enhancer module, we
performed Site Clustering Over Random Expectation (SCORE) (Rebeiz et al.,
2002) to determine whether this collection of proneural and bHLH-R sites was
statistically significant. Extending the SCORE approach to search for co-
clusters of two or more sites, we performed monte-carlo simulations of
randomized data sets to see how often observed cluster-frequencies could
occur by random chance. In SCORE analysis, the distinction between high
and low purity clusters is made: high purity clusters represent a cluster that is
so improbable to occur by random chance that there are likely to be zero or
very few randomly occurring clusters in the genome. Lower purity clusters may
be extremely likely to occur by random chance, but the SCORE approach dictates that one should look for cluster types that have frequencies that are at least two-fold over random expectations (50% purity). Only two regions showed high purity (≥95%) in our SCORE analysis of proneural + bHLH-R sites (Table 4.S2). Neither of these clusters exhibited strong conservation when orthologous regions were searched in *D. pseudoobscura* and *D. virilis*. However, *phyl* did show up in a cluster bin that was significant (*p* = .003 in 1000 simulations), and relatively pure (77%). Clusters in this bin contain three proneural sites, and two predicted bHLH-repressor sites in <400 bp. Of the eight clusters found in this bin, ~2 are expected to occur by random chance (random average = 1.8). That the genomic content of eight clusters of this class was reproduced only 3 times in 1000 random trials shows that the observed frequency of this type of cluster of sites is quite improbable due to random chance. Recent studies have pointed towards sequence conservation as an effective means of narrowing down transcription factor binding site clusters identified computationally (Berman et al., 2004). Using the criterion of conservation between *D. melanogaster*, *D. pseudoobscura*, and *D. virilis*, only one other cluster in the *phyl* bin showed a similar degree of sequence constraint. This cluster, residing in the first intron of *nervy* (Fig 4.5A) contains three proneurals and two bHLH-R sites spanning 343 bases in *D. melanogaster*. The intron containing this cluster also harbors an inordinate
number of proneural binding sites, a feature that is statistically significant as assessed by the SCORE technique (the region hit by SCORE on proneural binding sites alone does not overlap that hit by the proneural + bHLH-repressor combination). *nervy* is the *Drosophila* homolog of the mammalian oncogene ETO, and is a transcription factor known to be expressed in the SOP (Wildonger and Mann, 2005). Given our success with the *phyl* cluster, we investigated the potential SOP module of *nervy* that could implement an identical logic to *phyl*.

**The *nervy* cluster drives expression in SOPs**

Placing a 782 bp fragment from *nervy* into a GFP reporter construct, we observe that this fragment is sufficient to drive GFP expression in the wing margin triple row SOPs (Fig 4.6A), a location previously demonstrated to express nervy protein (Wildonger and Mann, 2005). This fragment also drives expression throughout the adepithelial adult muscle precursors, in a pattern resembling *twist* (Currie and Bate, 1991). By crossing to *phyl* RFP, we determined that this adepithelial expression does not coincide with notum macrochaete SOPs (not shown). This pattern may be the result of the addition of incomplete regulatory sequences driving spurious expression in this region, which does not normally express *nvv*. Marking the wing margin with an antibody to Hindsight, which labels SOPs (Pickup et al., 2002), we see that
GFP expression coincides with wing margin SOPs (Fig. 4.6B-B”). Consistent with our prediction of that repressor sites are required to repress non-SOP expression in this module, we observe robust derepression of GFP expression in the wing margin when bHLH-R sites are mutated in the 782 bp fragment (Fig. 4.6C-C”). This result suggests that \textit{nervy} is also a target of \textit{E(spl)} bHLH-R’s in proneural clusters.

\textbf{Discussion}

In this report, we have presented a new theoretical perspective of Notch mediated lateral inhibition in proneural clusters. For over a decade, the models have dictated that \textit{E(spl)} bHLH-R genes carry out the sole task of down-regulating proneural bHLH-activators (Jiménez and Ish-Horowicz, 1997). However, several pieces of evidence contradict this conclusion. First, overexpression of \textit{scute} can induce ectopic bristles (Rodriguez et al., 1990), but at nowhere near the levels caused by loss of function mutations removing the \textit{Enhancer of split} complex (Heitzler et al., 1996). Second, when the best-characterized bHLH-R site in the \textit{achaete-scute} complex is mutated in the context of a rescue construct, a failure of lateral inhibition is not observed (Van Doren et al., 1994). Instead, this rescue construct results in ectopic bristles in a pattern similar to \textit{hairy}, a bHLH-R that is not Notch regulated. Van Doren and colleagues convincingly validated this bHLH-R site as a bonafide \textit{hairy}
target. Needless to say, there must be some contribution of $E(spl)$ mediated repression of proneural genes. In mutations of the $E(spl)$ complex, proneural genes are upregulated. However, no evidence that we are aware of demonstrates that this regulatory loop is direct.

Another competing model in the action of $E(spl)$ bHLH-R genes is the heterodimerization of these factors with proneural bHLH-activators, and the resulting repression of proneural target genes through proneural binding sites (Giagtzoglou et al., 2003). However, a simple interpretation of this model would pose that the phyl enhancer would be effectively repressed in non-SOPs when the bHLH-R sites are mutated. Our data on two separate in vivo targets of $E(spl)$ must be reconciled within any model that incorporates this off-the-DNA action of bHLH-repressors.

**A third code in the proneural cluster**

We present in this study a new code for the proneural cluster. The two other proposed codes use proneural activators in different ways. Notch targets use low levels of proneurals in non-SOPs to synergize with Su(H), while total-cluster genes use a proneural only code to achieve ubiquitous expression throughout the cluster (Reeves and Posakony, 2005). In this report, our code uses proneurals in the setting where their levels are highest: the SOP. Still, it is remarkable that upon removal of repressor binding sites, we see such
strong de-repression of these enhancers in non-SOPs cells. Many models might leave SOP expression to control of the proneurals by fine-tuning a proneural only code to respond to only the highest levels of proneurals. However, in the case of the *phylopod* enhancer, we see that a different strategy is employed. We believe that this new code for neural precursors may help us understand the lingering questions of lateral inhibition.

**Complex regulation of *phyl* during bristle development**

Although the *phyl* enhancer identified in this study recapitulates the endogenous pattern of *phyl* expression in both wildtype and proneural mutant backgrounds, an upstream enhancer active in SOP cells has also been identified (Pi et al., 2004). Interestingly, this enhancer also contains proneural binding sites, that upon mutation fails to inactivate the enhancer. Although this suggests a more complex scheme for activation of these enhancers, the residual activity also leads us to believe that both regulatory regions are relevant. When one identifies a region based on binding sites alone, the possibility is raised that observed reporter expression is artifactual, due to the high density of binding sites placed in front of the core promoter. However, when these excellent matches are mutated, and residual expression is observed, the conclusion is easily made that additional unknown factors are responsible for activation, abating ones concerns about artifactual activation.
One possible reason that *phyl* contains two SOP enhancers is its dual role in bristle development. Although its mechanism of action in bristle development is poorly understood, *phyl* is required for both the SOP cell fate, and also Notch mediated cell fate decisions between sister cells when the SOP divides (Pi et al., 2001). Therefore, one enhancer region could be used to stabilize SOP commitment, while the other enhancer is required to fulfill the need for *phyl* later on in bristle development. We propose that the intron *phyl* enhancer is used to secure the SOP fate, since it is downstream of the E(spl) bHLH-repressors.

**Logical implementation of a proneural + bHLH-repressor code**

In this report, we have described two enhancers that are turned on in the SOP in response to the proneural proteins, and which are repressed in the non-SOP as the end result of Notch signaling. These findings add to the growing picture of how SOP selection works. In the case of *phyl*, SOP expression is a requirement: *phyl* mutants lack SOPs, as measured by the loss of expression of several SOP markers (Pi et al., 2001). Conversely, *phyl* overexpression leads to increased bristle density (Pi et al., 2001). The logic of these phenotypes is paralleled by the proneural + bHLH-repressor code implemented by the *phyl* enhancer. Proneural expression is elevated to its highest levels within the SOP, and these activators carry out the task of
establishing the SOP fate. However, Notch mediated activation of bHLH-repressors is required to turn off genes that would normally come on in the non-SOP cells. Again, it is logical that phyl would be turned off in non-SOP cells, because high levels of phyl expression could compromise the selection of a single SOP.

In contrast to the central role that phyl plays in the SOP fate, nvy appears to play a more peripheral role (Wildonger and Mann, 2005). nervy mutants have a normal complement of bristle precursors. However, in nervy clones, there is a slight bias for the SOP to be selected in a nervy+ territory, rather than in the mutant tissue. Additionally, nervy overexpression was extremely effective in causing bristle loss by extinguishing SOP fates. This phenotype was caused only when nervy is broadly expressed, and not when nervy is overexpressed in the SOP (Wildonger and Mann, 2005). However complicated the role of nervy in SOP selection might be, it is still the case that nervy expression outside of the SOP is deleterious for the SOP fate, and thus it makes sense that nervy would use bHLH-repressors to achieve SOP specific expression in the wing margin.

Use of codes in bioinformatics searches

As we learn more about the transcriptional networks that drive development, we can apply our accumulated knowledge with greater
confidence to bioinformatic and genomic methods. In this report we used a hypothetical code to identify new regulatory regions. The power of a code is that it can narrow down the search to find only the modules of interest. For example, in our previous search of the *Drosophila* genome for targets of the Notch pathway (Rebeiz et al., 2002), target genes generated from the study could be expressed in any of the dozens of settings where Notch signaling can be used in the fly. However, if one were to narrow down the search to Su(H) binding sites and proneural binding sites, this list would be predicted to enrich for lateral inhibitory Notch targets. When performing multi-factor SCORE on proneural+Su(H) binding sites, no significant clusters of binding sites are found (M. Rebeiz, unpublished observations). In effect, the specificity is turned up, but so is the noise. However, if one were to filter the list of clusters of Su(H) + proneural binding sites by conservation, one would find that many of known clusters exhibit remarkable conservation (Castro et al., 2005; Nellesen et al., 1999). The conservation of binding site clusters has been shown to be a good positive indicator of functional sequences that display enhancer activity (Berman et al., 2004). We believe that genomic approaches will greatly improve once high throughput conservation analysis of binding sites can be combined with genomic methods such as SCORE. The thinking behind the SCORE technique is that a cluster so densely packed with binding sites that is unlikely to occur by random chance indicates a regulatory function. By
narrowing down the list of binding sites in the genome to the conserved list, perhaps the combinations of bHLH-repressors and proneural sites, or Su(H) + proneural sites that are less impressive will also turn out to be statistically significant, allowing us to find with more ease the enhancers that implement new codes to orchestrate the transcriptional regulatory network driving development of the organism.
Figure 4.1. A hypothetical dual input model for an SOP enhancer module. In the Notch responsive non-SOP, an enhancer with binding sites for proneural (pn) bHLH-activators, and Suppressor of Hairless [Su(H)] turn on bHLH-repressors of the Enhancer of split [E(spl)] complex. These transcriptional repressors bind to and repress an SOP gene enhancer containing binding sites for both proneurals and bHLH-repressors. In the SOP, the absence of Notch signaling maintains the repressive state of E(spl) genes, allowing proneurals to bind to this same enhancer, and activate transcription, unchallenged by E(spl). The red “X” signifies that the gene is transcriptionally silenced.
Figure 4.2. A cluster of proneural and bHLH-repressor sites in the intron of *phyllopod*. (A) schematic view of proneural and bHLH-repressor sites across the phyllopod genomic region. Blue diamonds represent matches to the proneural consensus RCAGSTG. Red boxes are matches to the bHLH-repressor consensus GGCACGYGHY. Bar denotes limits of a 662 bp GFP reporter construct centered on Proneural and bHLH-repressor sites. (B) Schematic alignment of cluster of sites within the 662 bp fragment across nine *Drosophila* species. Note the fairly stringent spacing of elements. Site numbering in this schematic (P1-P3 and R1-R2) matches site numbering in (C) and Figure 4.3. (C) Site alignment of proneural and bHLH-repressor sites across nine *Drosophila* species. “.” Represents a large gap in the alignment. “-“ represents a single nucleotide gap. “//” represents excess sequence (>10 bp) that has been removed from the alignment. “*” represents a nucleotide that is conserved in all species tested. Bold letters are matches to binding site consensuses. Red letters are bases that deviate from these binding site definitions. Species abbreviations: ana = *D. ananassae*; ere = *D. erecta*; gri = *D. grimshawii*; mel = *D. melanogaster*; moj = *D. mojavensis*; pse = *D. pseudoobscura*; sec = *D. sechellia*; wil = *D. willistoni*; vir = *D. virilis*.  

**Figure 4.2.** A cluster of proneural and bHLH-repressor sites in the intron of *phyllopod*. (A) schematic view of proneural and bHLH-repressor sites across the phyllopod genomic region. Blue diamonds represent matches to the proneural consensus RCAGSTG. Red boxes are matches to the bHLH-repressor consensus GGCACGYGHY. Bar denotes limits of a 662 bp GFP reporter construct centered on Proneural and bHLH-repressor sites. (B) Schematic alignment of cluster of sites within the 662 bp fragment across nine *Drosophila* species. Note the fairly stringent spacing of elements. Site numbering in this schematic (P1-P3 and R1-R2) matches site numbering in (C) and Figure 4.3. (C) Site alignment of proneural and bHLH-repressor sites across nine *Drosophila* species. “.” Represents a large gap in the alignment. “-“ represents a single nucleotide gap. “//” represents excess sequence (>10 bp) that has been removed from the alignment. “*” represents a nucleotide that is conserved in all species tested. Bold letters are matches to binding site consensuses. Red letters are bases that deviate from these binding site definitions. Species abbreviations: ana = *D. ananassae*; ere = *D. erecta*; gri = *D. grimshawii*; mel = *D. melanogaster*; moj = *D. mojavensis*; pse = *D. pseudoobscura*; sec = *D. sechellia*; wil = *D. willistoni*; vir = *D. virilis*. 

**Figure 4.2.** A cluster of proneural and bHLH-repressor sites in the intron of *phyllopod*. (A) schematic view of proneural and bHLH-repressor sites across the phyllopod genomic region. Blue diamonds represent matches to the proneural consensus RCAGSTG. Red boxes are matches to the bHLH-repressor consensus GGCACGYGHY. Bar denotes limits of a 662 bp GFP reporter construct centered on Proneural and bHLH-repressor sites. (B) Schematic alignment of cluster of sites within the 662 bp fragment across nine *Drosophila* species. Note the fairly stringent spacing of elements. Site numbering in this schematic (P1-P3 and R1-R2) matches site numbering in (C) and Figure 4.3. (C) Site alignment of proneural and bHLH-repressor sites across nine *Drosophila* species. “.” Represents a large gap in the alignment. “-“ represents a single nucleotide gap. “//” represents excess sequence (>10 bp) that has been removed from the alignment. “*” represents a nucleotide that is conserved in all species tested. Bold letters are matches to binding site consensuses. Red letters are bases that deviate from these binding site definitions. Species abbreviations: ana = *D. ananassae*; ere = *D. erecta*; gri = *D. grimshawii*; mel = *D. melanogaster*; moj = *D. mojavensis*; pse = *D. pseudoobscura*; sec = *D. sechellia*; wil = *D. willistoni*; vir = *D. virilis*. 

**Figure 4.2.** A cluster of proneural and bHLH-repressor sites in the intron of *phyllopod*. (A) schematic view of proneural and bHLH-repressor sites across the phyllopod genomic region. Blue diamonds represent matches to the proneural consensus RCAGSTG. Red boxes are matches to the bHLH-repressor consensus GGCACGYGHY. Bar denotes limits of a 662 bp GFP reporter construct centered on Proneural and bHLH-repressor sites. (B) Schematic alignment of cluster of sites within the 662 bp fragment across nine *Drosophila* species. Note the fairly stringent spacing of elements. Site numbering in this schematic (P1-P3 and R1-R2) matches site numbering in (C) and Figure 4.3. (C) Site alignment of proneural and bHLH-repressor sites across nine *Drosophila* species. “.” Represents a large gap in the alignment. “-“ represents a single nucleotide gap. “//” represents excess sequence (>10 bp) that has been removed from the alignment. “*” represents a nucleotide that is conserved in all species tested. Bold letters are matches to binding site consensuses. Red letters are bases that deviate from these binding site definitions. Species abbreviations: ana = *D. ananassae*; ere = *D. erecta*; gri = *D. grimshawii*; mel = *D. melanogaster*; moj = *D. mojavensis*; pse = *D. pseudoobscura*; sec = *D. sechellia*; wil = *D. willistoni*; vir = *D. virilis*. 

**Figure 4.2.** A cluster of proneural and bHLH-repressor sites in the intron of *phyllopod*. (A) schematic view of proneural and bHLH-repressor sites across the phyllopod genomic region. Blue diamonds represent matches to the proneural consensus RCAGSTG. Red boxes are matches to the bHLH-repressor consensus GGCACGYGHY. Bar denotes limits of a 662 bp GFP reporter construct centered on Proneural and bHLH-repressor sites. (B) Schematic alignment of cluster of sites within the 662 bp fragment across nine *Drosophila* species. Note the fairly stringent spacing of elements. Site numbering in this schematic (P1-P3 and R1-R2) matches site numbering in (C) and Figure 4.3. (C) Site alignment of proneural and bHLH-repressor sites across nine *Drosophila* species. “.” Represents a large gap in the alignment. “-“ represents a single nucleotide gap. “//” represents excess sequence (>10 bp) that has been removed from the alignment. “*” represents a nucleotide that is conserved in all species tested. Bold letters are matches to binding site consensuses. Red letters are bases that deviate from these binding site definitions. Species abbreviations: ana = *D. ananassae*; ere = *D. erecta*; gri = *D. grimshawii*; mel = *D. melanogaster*; moj = *D. mojavensis*; pse = *D. pseudoobscura*; sec = *D. sechellia*; wil = *D. willistoni*; vir = *D. virilis*. 

**Figure 4.2.** A cluster of proneural and bHLH-repressor sites in the intron of *phyllopod*. (A) schematic view of proneural and bHLH-repressor sites across the phyllopod genomic region. Blue diamonds represent matches to the proneural consensus RCAGSTG. Red boxes are matches to the bHLH-repressor consensus GGCACGYGHY. Bar denotes limits of a 662 bp GFP reporter construct centered on Proneural and bHLH-repressor sites. (B) Schematic alignment of cluster of sites within the 662 bp fragment across nine *Drosophila* species. Note the fairly stringent spacing of elements. Site numbering in this schematic (P1-P3 and R1-R2) matches site numbering in (C) and Figure 4.3. (C) Site alignment of proneural and bHLH-repressor sites across nine *Drosophila* species. “.” Represents a large gap in the alignment. “-“ represents a single nucleotide gap. “//” represents excess sequence (>10 bp) that has been removed from the alignment. “*” represents a nucleotide that is conserved in all species tested. Bold letters are matches to binding site consensuses. Red letters are bases that deviate from these binding site definitions. Species abbreviations: ana = *D. ananassae*; ere = *D. erecta*; gri = *D. grimshawii*; mel = *D. melanogaster*; moj = *D. mojavensis*; pse = *D. pseudoobscura*; sec = *D. sechellia*; wil = *D. willistoni*; vir = *D. virilis*.
Figure 4.3. Binding of proneurals and bHLH-repressors to E-Boxes within the phylopod intron. (A) Shifts of proneural binding sites by achaete-daughterless heterodimers. A positive control probe from $E(spl)$ m4 cannot bind GST-Achaete alone (“ac” lane), but does bind to GST-Daughterless homodimers (“da” lane). This same probe shows a tight binding to Achaete/Daughterless heterodimers (“ac + da”). The phy1 P1-3 probes (see schematic in Fig 4.2) bind the heterodimer as well. The phy1 P1 mutant (mut) is a two base pair mutation in phy1 P1 that abolishes proneural binding. (B) Shifts of bHLH-R sites by a truncated form of $E(spl)$ m7. The high affinity bHLH-repressor binding site found upstream of achaete (ac h E1) was used as a positive control probe. A three base pair mutation in the phy1 R2 probe abolishes binding of m7 truncated.
Figure 4.4. Expression of *phyloped* intron reporters in 3rd instar wing imaginal discs. (A-C) shows wildtype RFP reporter compared to a wildtype GFP reporter (A’), a bHLH-repressor mutant GFP reporter (B’), and a proneural mutant GFP reporter (C’). (A”-C”’) merged GFP/RFP images of discs. The bHLH-repressor mutant (B’) shows expansion of nearly every wing disc cluster, in comparison to the wildtype reporter. Likewise, the proneural mutant (C’) is weakened in every position, and shows a general de-repression in the wing pouch region. Bristle position abbreviations in B’’: scutellar (sc), dorsocentral (dc), notopleural(np), wing margin (wm), dorsal radius (dR).
Figure 4.5 A cluster of proneural and bHLH-repressor sites in the *nervy* locus. (A) Schematic of the *nervy* gene region. Boxed area denotes region shown in (B). Bar show the limits of the 782 bp fragment used in a reporter assay. (B) Phylogenetic footprinting of proneural and bHLH-R site cluster. Positions of proneural and bHLH-R sites from the orthologous region of *D. pseudoobscura* and *D. virilis* are shown. Legends present consensuses used for proneural binding sites and bHLH-repressor sites in IUPAC code.
Figure 4.6 An SOP enhancer of \textit{nvy} restricted by bHLH-R binding sites. (A) expression of GFP reporter under control of the 782 bp \textit{nvy} region, containing three proneural binding sites, and two bHLH-R sites. Expression is seen in wing margin bristle SOPs, as well the ectopic activation in adult muscle precursors. (B,C) Wing margin SOPs marked by the Hindsight antibody. (B') GFP is expression of the wildtype \textit{nvy} enhancer is restricted to SOPs of the wing margin, however, when bHLH-R sites are mutated (C'), expression outside of SOP cells is observed. (B''-C'') merged images.
Figure 4.S1. Survey of multiple wildtype *phyt* GFP lines. GFP expression of 4 wildtype lines (A'-D') are compared to the expression of the single wildtype RFP line used in this study (A-D). GFP and RFP were imaged simultaneously in the same tissue (merge, A''-D''). Note the generally robust GFP expression that coincides well with RFP expression. Insertion line number is listed at the bottom of (A''-D'').
Figure 4.S2. Survey of multiple repressor mutant *phyt* GFP lines. GFP expression of 4 repressor mutant lines (A'-D') are compared to the expression of the single wildtype RFP line used in this study (A-D). GFP and RFP were imaged simultaneously in the same tissue (merge, A''-D''). Note the extremely strong expression of GFP that is expanded relative to RFP in nearly all clusters. Insertion line number is listed at the bottom of A''-D''.
Figure 4.S3. Survey of multiple proneural mutant *phyli* GFP lines. GFP expression of 4 proneural mutant lines (A’-D’) are compared to the expression of the single wildtype RFP line used in this study (A-D). GFP and RFP were imaged simultaneously in the same tissue (merge, A’’-D’’). Note the weakened expression of all GFP lines compared to wildtype RFP, and the wildtype GFP lines presented in Fig. 4.S1 (taken at identical gain). Also, wing pouch de-repression is present in all lines, but evident to different degrees. Insertion line number is listed in A’’-D’’.
Figure 4.S4 Expression of phyl reporter in the absence of achaete and scute. The phyl wildtype reporter was crossed into a sc$^{10-1}$ background. (A) Wing disc showing GFP loss in all proneural territories where achaete and scute are required. The cluster of GFP cells in the dorsal radius is a sensory position dependent on the proneural atonal. (B) In the eye imaginal disc of a sc$^{10-1}$ fly, GFP is still present, as most sensory cells depend on atonal.
Table 4.S1 Primers used in this study to create EMSA probes and reporter constructs. Lowercase letters in primer sequences mark *Kpn* I and *Bgl*II sites that were engineered into the PCR product to clone into the hStinger GFP reporter vector (Barolo et al., 2000).

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ac h E1</td>
<td>GGCAGCCGGCCAGCGGACAGGG</td>
</tr>
<tr>
<td>phyl R1</td>
<td>TCGCGCATCGCGGCTGCTAATGTT</td>
</tr>
<tr>
<td>phyl R2</td>
<td>TGGATAGGGCGGGCTGCTCTTGCC</td>
</tr>
<tr>
<td>phyl R2 mut</td>
<td>TGGATAGGGAGAGGGGCTCTTGCC</td>
</tr>
<tr>
<td>m4 E1</td>
<td>CAAAAACGGCCAGGTGCTTTTC</td>
</tr>
<tr>
<td>phyl P1</td>
<td>TTTCTCGGCAGCTGCTCGCCAT</td>
</tr>
<tr>
<td>phyl P1 mut</td>
<td>TTTCTCGCCAGCGGCTGCGGCA</td>
</tr>
<tr>
<td>phyl P2</td>
<td>TTTTGGCCAGCTGCTACGGCTG</td>
</tr>
<tr>
<td>phyl P3</td>
<td>TTTTGCTGAGCAAGTTGCA</td>
</tr>
</tbody>
</table>

**Reporter Constructs**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>nvy-F</td>
<td>gaagatcttcATGCACTACATCTCTGCTGTTGG</td>
</tr>
<tr>
<td>nvy-R</td>
<td>ggggtacccccCCCTCGATTCACCACCTTCTATTT</td>
</tr>
<tr>
<td>phyl-F</td>
<td>ggggtacccccTACGCCTATGCACTTAGTGATGCG</td>
</tr>
<tr>
<td>phyl-R</td>
<td>gaagatcttcTTTGCACCTGCGCTGCAATTTGCA</td>
</tr>
</tbody>
</table>

**Mutagenesis Primers**

<table>
<thead>
<tr>
<th>Name</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>phyl R2 mut-F</td>
<td>TCGGCAGAGGCCCCCTCCTCCCTATCCATAT</td>
</tr>
<tr>
<td>phyl R2 mut-R</td>
<td>ATATGGATAGGGGAGGGGTGCTTGGCCGA</td>
</tr>
<tr>
<td>phyl R1 mut-F</td>
<td>ACGACCATTAGGCCCTCCTCATGCGGACAG</td>
</tr>
<tr>
<td>phyl R1 mut-R</td>
<td>CTGTGCGCATGAGGGGCTCTATGCGT</td>
</tr>
<tr>
<td>phyl P2/P3 mut-F</td>
<td>TCGTGCCCCTCCAGCAAAAAACTAAGCGCCACGCGTGATAAGCGGGCCCA</td>
</tr>
<tr>
<td>phyl P2/P3 mut-R</td>
<td>TGGGCCGCCCTTAGCGCTGCTTTGGCTGAGGACGCTG</td>
</tr>
<tr>
<td>phyl P1-mut-F</td>
<td>GCGATCATCCGCTTCGCGGAGAAATT</td>
</tr>
<tr>
<td>phyl P1-mut-R</td>
<td>TAAATTTCTCGCGGAAAGGCTCGGCAGATGCG</td>
</tr>
<tr>
<td>nvy R1 mut-F</td>
<td>CATACTGCTCGCCAGCTGCTTTATAT</td>
</tr>
<tr>
<td>nvy R1 mut-R</td>
<td>AACACATGCTGAAAGGCTCGACACTGATG</td>
</tr>
<tr>
<td>nvy R2 mut-F</td>
<td>CTGGAATAGCGCTCCCTTTCTGCGAAAGGCTG</td>
</tr>
<tr>
<td>nvy R2 mut-R</td>
<td>CACGCCTTGGCAGAAAGGACGCTATTCAAG</td>
</tr>
</tbody>
</table>
Table 4.S2. Clusters of proneural + bHLH-repressor sites identified by SCORE in the same significance bin as *phyl* (2 bHLH-repressors and 3 proneurals in < 400 bp).

<table>
<thead>
<tr>
<th>Gene Name</th>
<th>CG #</th>
<th>Relative position of cluster</th>
</tr>
</thead>
<tbody>
<tr>
<td>CG17508</td>
<td>CG17508</td>
<td>(+)3028</td>
</tr>
<tr>
<td><em>phyl</em></td>
<td>CG10108</td>
<td>Intron 1</td>
</tr>
<tr>
<td><em>nvy</em></td>
<td>CG3385</td>
<td>Intron 1</td>
</tr>
<tr>
<td><em>rl</em></td>
<td>CG12559</td>
<td>(-)35568</td>
</tr>
<tr>
<td>CG15766</td>
<td>CG15766</td>
<td>(+)836</td>
</tr>
<tr>
<td>CG15579</td>
<td>CG15579</td>
<td>(+)3670</td>
</tr>
<tr>
<td>CG6247</td>
<td>CG6247</td>
<td>(-)4138</td>
</tr>
</tbody>
</table>
References


CHAPTER 5
AN ANCIENT TRANSCRIPTIONAL REGULATORY LINKAGE
An ancient transcriptional regulatory linkage

Mark Rebeiz, Tammie Stone, James W. Posakony*

Division of Biological Sciences, Section of Cell and Developmental Biology, University of California San Diego, La Jolla, CA 92039-0349, USA

Received for publication 2 February 2005, revised 8 March 2005, accepted 8 March 2005
Available online 11 April 2005

Abstract

Changes in gene regulatory networks are a major engine for creating developmental novelty during evolution. Conversely, regulatory linkages that survive for very long evolutionary periods might be characteristic of ancient and abstract functions of fundamental utility to all metazoans. The proneural genes, which encode a distinctive family of basic helix-loop-helix (bHLH) transcriptional activators, act to promote neural cell fate in the ectoderm of diverse species. Here we report that these genes have been associated for at least 600–700 million years—since before the cnidarian/bilaterian divergence—with a high-affinity binding site for Hairy/Enhancer of split (Hes) repressor proteins. We suggest that the systematic identification of such ancient and conserved connections will be a powerful means of uncovering the primordial functions of transcription factors and signaling systems.

© 2005 Elsevier Inc. All rights reserved.

Keywords: Proneural genes; Achaete-Scute; Atonal; bHLH repressors; Binding site; Transcriptional repression; Gene phylogeny; Phylogenetic footprinting; Cnidarians; Metazoan evolution

Introduction

In both protostomes and deuterostomes, the specification of neural cell fates in the ectoderm depends on the proneural genes, which encode a distinctive family of basic helix-loop-helix (bHLH) transcriptional activators. The proneural genes belong to two major subfamilies, those related to the Drosophila genes achaete (ac) and scute (sc), and those related to Drosophila atonal (ato). That the proneural gene family originated deep in metazoan evolutionary history is indicated by the identification of clear homologs in cnidarians such as Hydra (Grens et al., 1995) and the jellyfish Podocoryne carnea (Muller et al., 2003; Seipel et al., 2004). Conservation of proneural gene function over very large evolutionary distances is demonstrated dramatically by the observations that Hydra CnAsh efficiently promotes sensory bristle formation in adult Drosophila (Grens et al., 1995) and that Drosophila ato can fully substitute for its homolog Atoh1 in the mouse (Wang et al., 2002).

Spatial and temporal patterns of proneural gene activity are shaped during development by a variety of negative regulators that are themselves members of the HLH protein family. In Drosophila, the pattern of sensory organs on the body surface is sculpted by the bHLH repressor Hairy, which binds directly to a 10-bp, high-affinity binding site upstream of ac and represses its expression (Ohsako et al., 1994; Van Doren et al., 1994). Specific mutation of this GGCACCGGAC motif in a genomic DNA transgene causes the appearance of ectopic sensory organs in the adult fly, a phenotype closely mimicking that conferred by loss-of-function mutations in the hairy gene. Individual neural precursor cells in Drosophila are specified by the process of “lateral inhibition”, mediated by the Notch (N) cell–cell signaling pathway, in which all but one cell in a “proneural cluster” of potential precursor cells are prevented from adopting this fate (reviewed in Lai, 2004). Here, proneural gene expression is antagonized by multiple bHLH repressors encoded by N-activated genes of the Enhancer of split Complex (Heitzler et al., 1996). Genetic studies indicate that mammalian proneural genes are likewise subject to negative regulation by bHLH repressors (reviewed in Davis and Turner, 2001; Ross et al., 2003). Thus, targeted disruption of
the mouse *Hairy and Enhancer of split homolog-1* (Hes-1) gene leads to elevated expression of the proneural gene *achaete-scute-like-1* (*Ascl1* or *Mash1*) and premature neurogenesis (Cau et al., 2000; Ishibashi et al., 1995).

Here we report that, in both protostomes and deuterostomes, proneural genes of both the *ac-sc* and *ato* subfamilies are widely associated with high-affinity binding sites for Hes-type bHLH repressors. Long-term evolutionary conservation of exact site sequences, coupled with a phylogenetic analysis of site distribution among proneural genes, provides strong evidence of inheritance of this binding site from a progenitor proneural gene in a metazoan ancestor that lived at least 600–700 million years ago (Mya). Consistent with this interpretation, we show that bHLH repressor proteins from a cnidarian, *Drosophila*, and rat have similar binding specificities in vitro. We suggest that ancient, conserved regulatory linkages of this kind are characteristic of abstract developmental functions of fundamental utility to all metazoans.

Materials and methods

**Motif searches**

A search criterion representing high-affinity binding sites for bHLH repressors of the Hairy/Enhancer of split (Hes) class was developed by merging the optimal specificity defined for *Drosophila* Hairy (GGCACGCGMC; Van Doren et al., 1994) with that defined for *Drosophila* Enhancer of split proteins (GGACGTGGY; Jennings et al., 1999), yielding GGCACGYGHY. Repressor binding sites were detected and visualized in genomic DNA sequence using the GenePalette software tool (Rebeiz and Posakony, 2004). Most genome sequences and annotations were retrieved from GenBank (Benson et al., 2004), while some newer genome sequence data (e.g., chicken, dog, zebrafish, pufferfish) were accessed through EnsEMBL. (Birney et al., 2004). Nearly all proneural proteins are encoded by a single open reading frame (ORF); in most cases, binding sites were detected by searching a 1500-bp window upstream of the ORF. Only in the case of *Caenorhabditis* *lin-32* genes, in which a repressor site appears ~300 bp downstream of the transcription unit, were sites outside of the 1500-bp window counted as present. Where only genomic DNA sequence traces were available (sea urchin, amphioxus, sea anemone, trematode, flatworm), proneural protein coding sequences were recovered by local TBLASTN, and traces were assembled into contigs using the CAP3 online contig assembly tool (Huang and Madden, 1999). ORFs were detected and verified using the ORF Finder utility at NCBI (http://www.ncbi.nlm.nih.gov/gorf/gorf.html). A gene was counted as not bearing a site only if 1500 bp upstream of the ORF was available for search. An exact match to the GGCACGYGHY binding site definition was required to call a site as present. Single-mismatch sites were scored only if the mismatch occurred in one of the four nucleotides flanking the required CACGYG core (see text and Fig. 1).

**Phylogenetic trees**

An alignment (available upon request) of the DNA sequences that encode the bHLH domains of 159 proneural genes was used to construct phylogenetic trees. Amino acid sequences were aligned using ClustalX (http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/Top.html); it was found that the bHLH domain is the only region that is alignable with confidence when all proneural proteins are considered. The ClustalX protein alignment was converted to a nucleotide alignment (Perl script available upon request), and the alignment was manually edited using the Se-Al program (http://evolve.zoo.ox.ac.uk/).

Trees were inferred by Bayesian likelihood analysis using MrBayes 3 (Ronquist and Huelsenbeck, 2003). Two independent runs of 10 million iterations were performed; for each the burn-in was 2 million cycles. The replicate trees are nearly identical, with only one branch (Am_ac) placed differently. The tree in Fig. 2 was rooted using the ASH and ATH/ATOH subfamilies as outgroups for each other. The tree in Supplementary Fig. S1 is the second replicate. The size of the data set precluded construction of a reliable tree using maximum parsimony or maximum likelihood.

**Gel mobility shift assays**

Electrophoretic mobility shift assays (EMSAs) were performed as previously described (Bailey and Posakony, 1995; Van Doren et al., 1994) using rat Hes-1 and *Drosophila* Hairy-truncated GST fusion proteins (Grbavec and Stifani, 1996; Van Doren et al., 1994). A plasmid construct that expresses a *Nematostella vectensis* GST-Hes-1 fusion protein (*Nv* Hes-1) was prepared as follows. The sequences encoding the N-terminal 100 aa of the *Nematostella* Hes-1 protein (P. Trowb, unpublished observations) were PCR-amplified using the following primers:

```
ggcatecGCGCCTCTCTTTTACCA
```

Lowercase letters denote BamHI and EcoRI sites used to clone the PCR fragment into pGEX-5X-2 (Amersham Pharmacia). Oligonucleotides used as probes in the EMSA:

```
Dm_ac GCCACGCGCAGCGACAGGG
Dm_ac_r GGCACGCGCAGCGACAGGG
Fr_ac GCTTACACCCACGATCCGCCG
Am_ato GCCCGAGCGACGAGCACTGAG
Ce_ato TGATTTTGCGCAGGTCATTG
```
Ce ato m TGATTTCGCCCCCTTTCTATGTT
Bf nge GTGCTGAGGGCAACCGGCTGCGGTC
Mn ato1 TAAAAACCGGACACGGCGACGTAT
Hs nge2 CGGCTCCTGGACCGCGACTCCAGG

Results

Widespread association of bilaterian proneural genes with binding sites for bHLH repressors

The Drosophila melanogaster proneural gene ac is subject to direct transcriptional regulation by the bHLH repressor protein Hairy, acting through a 10-bp high-affinity binding site, GGCACCGGAC (Figs. 1A–B). Evolutionary conservation of this regulatory linkage over perhaps 60 million years (My) is indicated by the presence of a clearly orthologous site within a 13-bp block of sequence identity upstream of the ac genes of multiple Drosophila species (Fig. 1B). We further observed that the ac-sc orthologs of the mosquitoes Anopheles gambiae (Fig. 1A) and Aedes aegypti (Dipterans, like Drosophila), the silk moth Bombyx mori (a Lepidoptera), the red flour beetle Tribolium castaneum (a Coleoptera), and the honeybee Apis mellifera (a Hymenoptera) also include in their proximal upstream sequences a similar 10-bp element predicted to be a high-affinity binding site (GGCACGGYHY) for bHLH repressor proteins of the Hairy/Enhancer of split (Hes) class (Jennings et al., 1999; Van Doren et al., 1994) (Fig. 1B). In all cases, the element is found within 1500 bp upstream of the translation start codon. In DNA with no base composition bias, such motifs are expected to appear randomly every 43,691 bp, so it is highly unlikely that the observed occurrences are due to chance. This depth of regulatory sequence conservation (approximately 270 My) (Burmeister et al., 1998) surpasses that which has been observed in insects to date (Erives and Levine, 2004; Markstein et al., 2004).

In attempting to trace the evolutionary origin of this transcription factor/target linkage, we observed that a predicted high-affinity bHLH repressor binding site is also present within the same 1500-bp window upstream of an ac-sc homolog (ASH) in all deuterostome species examined (Figs. 1A and B), including human, mouse, zebrafish, sea urchin, and ascidian. The ascidian element has exactly the same sequence (GGCACGGCAG) as the sites upstream of Drosophila ac and the Anopheles ac-sc ortholog. Mammals and zebrafish share an element (GGCACGGCAGC) that is only one base different from this, while the site in pufferfish and sea urchin is a different one-base variant, GGCACGGTCT. All of these motifs (Fig. 1B) are consistent with high-affinity binding by bHLH repressors (Jennings et al., 1999; Van Doren et al., 1994), and their presence upstream of ASH genes in both protostomes and deuterostomes suggests that this regulatory linkage was established in a bilaterian ancestor.

This in turn raises the question of whether proneural genes of the ato subfamily might be similarly regulated. We find that a predicted high-affinity bHLH repressor binding site (GGCACGTGCT) is precisely conserved in association with all insect ato genes and that a similar site (GGCACGGCC) is present upstream of an ato-homologous (ATH/ATOH) gene in all deuterostome species for which data are available, including mammals, Xenopus, and fish (Figs. 1A and C). The 10-bp element GCCACCGGCT, also consistent with high-affinity binding by bHLH repressors, is present 300 bp downstream of the C. elegans ato homolog lin-32 and is conserved in Caenorhabditis briggsae (Fig. 1C). It is generally accepted that the ac-sc and ato subfamilies of proneural bHLH genes evolved from a common ancestor. The very broad phylogenetic distribution of potential bHLH repressor binding sites in association with genes of both subfamilies suggests that a bHLH repressor site was already associated with that deep ancestor.

We have also detected conservation of bHLH repressor binding sites in specific paralogous gene families of both the ac-sc and ato classes, asease (ase) is an insect-specific paralog of the ac-sc subfamily (Fig. 1A). The 10-bp element GCCACCGGAC is precisely conserved upstream of this gene from Drosophila to Anopheles and is present with a single substitution in the honeybee and flour beetle ase genes (Fig. 1D). We note that bHLH repressor sites are located in the 5' UTRs of a minority of paralinear genes (e.g., Drosophila ase). A strong precedent for the functionality of transcriptional repressor binding sites in the 5' UTRs of target genes is provided by the factor REST/NR3C1, as Schoenherr et al. (1996). Two families of ato paralogs, one in insects and one in deuterostomes, also conserve a predicted high-affinity bHLH repressor binding site, cousin of atonal (cato) (Goulding et al., 2000a) is associated with the element GCCACGGCGTC in all species of Drosophila examined (Drosophila ananassae has a one-base substitution); the site GCCACGTGCT is found in the mosquito (Figs. 1A and E), neurogenin genes from amphioxus, fish, amphibians, chicken, and mammals are likewise associated with various versions of a bHLH repressor site; in mammals, the 10-bp element (GCCACCGGAC) is part of a stable 15-bp block (Figs. 1A and F). Thus, even following relatively ancient proneural gene duplication events, a regulatory linkage with bHLH repressors appears to be retained in certain paralogs.

A global survey of the putative bHLH repressor sites we have observed in association with proneural genes reveals several underlying patterns and generalizations (see Supplementary Table S1). In almost all cases (60/62), only a single sequence element fitting our search criterion (the GGCACCGYHY binding site definition) is found in a given gene. There is no significant orientation bias overall, but there is a very strong tendency to conserve site orientation in orthologous genes, sometimes over large evolutionary distances. Thus, all ASH genes except that of the sea urchin
have the bHLH repressor site in the forward orientation; the motif is forward in all insect *ato* genes, while the reverse orientation appears in all *neurogenin* genes. There is a moderate positional bias; the site is located within the 300 bp proximal to the translation start site in 30/62 cases, though a 1500-bp window was allowed. Finally, very clear biases are observed at the degenerate positions of the 10-bp binding site, such that only seven of the 12 versions of the site permitted by our search criterion actually appear in the dataset (Figs. 1B–F). Within individual prooneural gene subfamilies, there is a very strong preference for one or the other core hexamer in the site (CACCAC vs. CAGTGC). In every site found upstream of an *ASH, atox, or neurogenin* gene, the core hexamer is always CACCAC (Figs. 1B, D, and F). In the *ATH/ATOH* subfamily, insects utilize the CAGTGC hexamer, while the chordate sites all have the
CACCG core (Fig. 1C). Overall, only 10 of the 64 sites have a CACGTG core hexamer. A sequence logo (Crooks et al., 2004; Schneider and Stephens, 1990) constructed from the entire dataset is shown in Fig. 1G.

An important question in evaluating the significance of these site occurrences is whether relaxation of the search criteria (the high-affinity binding site definition GGGAC-GYGHY) would yield a large number of additional occurrences. To preserve recognition by bHLH repressors, we constrained the core hexamer of the 10-bp site to CAGGYG, but permitted a single-base substitution at one of the four remaining positions (the two 5'-most and the two 3'-most; respectively, 1/2 and 9/10 in Fig. 1G). In the entire set of 148 proneural genes examined in this study, we observed 55 additional sites fitting the relaxed search criteria; i.e., fewer than those meeting the stringent criteria. The significance of this finding may be appreciated by considering that a total of 100 different 10-mers meet the relaxed criteria, yet only seven of these variants (Figs. 1B–F) account for more than half of the observed occurrences. Randomly, the seven stringent sites would be expected to account for a total of only eight occurrences; instead they account for 64. We conclude that bilaterian proneural genes are associated highly non-randomly with 10-bp elements of the sequence distribution shown in Fig. 1G.
Phylogenetic distribution of bHLH repressor binding sites associated with proneural genes

To examine phylogenetic relationships between genes that contain the bHLH repressor site and those that do not, we performed Bayesian likelihood analyses using the bHLH domain-encoding sequences of the proneural genes (Ronquist and Huelsenbeck, 2003). The resulting tree strongly supports the existence of at least three distinct bHLH subfamilies (Fig. 2). Clear patterns suggestive of repressor binding site inheritance are immediately apparent (Fig. 2, Fig. S1). Among vertebrate ato homologs, all genes of the *Atoh1 (Ath1)* clade include a site, while no sites are found associated with genes of the *Atoh7 (Ath3)* sister clade. Similarly, in the insect ato subfamily, all genes in the ato clade contain the site, while the paralogous *amor* clade (Goulding et al., 2000b; Huang et al., 2000) is devoid of sequences meeting the stringent criteria. The insect *ac-sc* subfamily also displays a consistent pattern of site retention. The *Drosophila achaete-scute Complex* contains three ASH genes: ac, sc, and lethal of scute, but this is atypical; most other insects (e.g., *A. gambiae, A. mellifera*, and *T. castaneum*) have a single ASH gene (Galant et al., 1998; Wheeler et al., 2003; Wolbeck and Simpson, 2002). Of the three *Drosophila* ASH genes, only ac has a binding site upstream, while the single ASH gene of the other insects is consistently associated with an upstream site. The Ascl1/2 and Ascl3/4 clades represent an extremely deep ASH gene duplication event; these paralogs are conserved from cnidarians to mammals. Yet the latter clade entirely lacks predicted high-affinity bHLH repressor sites, while at least the chordate *Ascl1* and *sea urchin* ASCl genes within the Ascl1/2 clade are associated with such sites.

The *neurogin* genes present an interesting picture. Genes of the *neurogin 2* clade are associated with a repressor binding site, while *neurogin 1* and 3 genes are not. In teleost species, *neurogin 1* and 2 subfamilies are not represented, and instead a single gene is found that groups into a fish-specific clade containing a repressor site. This single teleost *neurogin* gene has been suggested to represent the common ancestor of *neurogin 1* and 2, performing functions in fish that are delegated to separate genes in higher vertebrates (Andermann et al., 2002; Cornell and Eisen, 2002).

All of these instances are consistent with an inherited bHLH repressor site being retained by only one of the progeny of a gene duplication event. In a subfunctionalization model, for example, this could happen if only one of the duplicated genes retained the expression specificity (or specificities) that required or utilized Hes-mediated repression (Lynch and Force, 2000).

Several clades include genes with an element meeting only the relaxed search criteria described above (purple lettering in Fig. 2), but these occurrences show far less phylogenetic consistency than the stringent sites, and most are likely due to chance (1 expected every 5958 bp). In only two instances (Dr *neuroD* and Cp *neuroG*, Fig. 2) is a stringent site found within a clade of closely related genes that otherwise do not contain the site. This also is consistent with random expectation for the total sequence length searched (183 kb; four expected).

By making use of the NCBI trace sequence database (http://www.ncbi.nlm.nih.gov/Traces/), we have identified proneural gene homologs belonging to both the ASH and ATH/ATOH subfamilies in the genome of the sea anemone *N. vectensis*, a cnidarian. In one *Nematostella* homolog of *ato* (Nv *ath2*, Fig. 2), we find an element meeting our stringent site criteria 1200 bp upstream (Fig. 1C). While this observation needs to be supported by evidence from other cnidarian species, it is consistent with an ancient origin for the proneural gene/bHLH repressor connection.

The ASH and ATH/ATOH subfamilies appear to have been founded by an ancient duplication of a progenitor proneural gene (Fig. 2). Since at least two extant cnidarians (the sea anemone *N. vectensis* and the jellyfish *P. carnea*) have clearly recognizable homologs of both the ASH and ATH/ATOH types, it may be presumed that the divergence of the two subfamilies predates the divergence of the diploblast and bilaterian lineages. We suggest that the extensive association of protostome, deuterostome, and cnidarian proneural genes with putative high-affinity bHLH repressor binding sites indicates a common regulatory linkage that has been preserved from a deep metazoan ancestor that lived at least 600–700 Mya (Douvery et al., 2004; Peterson et al., 2004).

Evolutionary conservation of the DNA-binding specificities of Hes-type repressor proteins

An important implication of the above proposal is that bHLH repressor proteins from any species should be able to bind efficiently in vitro to the stringent (GGCACGYGHY) sites we have defined (Figs. 1B–G), regardless of their source or exact sequence. Fig. 3 shows that this expectation is indeed borne out. First, we observe strong sequence-specific binding by bHLH repressors from three distantly related species (*Nematostella, Drosophila*, and rat) to two sites that differ in two of the three degenerate positions (Fig. 3A). Moreover, the rat Hes-1 protein binds efficiently to examples of the six most prevalent site variants, taken from diverse species (Fig. 3B). These data validate the prediction that Hes-class bHLH repressor proteins have not substantially modified their DNA-binding specificity since the cnidarian/bilaterian divergence.

Discussion

A conserved transcriptional regulatory linkage

The most parsimonious interpretation of the results presented in this report, and the one that we favor, is that
a progenitor proneural gene in an early metazoan ancestor became subject to regulation by one or more Hes-class bHLH repressors via a high-affinity binding site fitting the specificity shown in Fig. 1G; and that this site has been maintained by selection in multiple orthologous and paralogous descendants of this progenitor for at least 600–700 My. We cannot strictly rule out the possibility that only the mode of regulation has been maintained and that the binding site itself has been replaced repeatedly in the course of animal evolution. However, several lines of evidence suggest that the more stringent interpretation is to be preferred.

First, the strong tendency of the site to persist, without apparent exception, in certain ancient clades of clearly homologous genes (e.g., the ac-sc, ato, and neurogenin homologs; Figs. 1 and 2) is more compatible with continued selection for an inherited site rather than frequent loss and reacquisition. The complete absence of the site from certain other clades of proneural genes (e.g., AtoH7, neurogenin 3, Ascl3/4; Fig. 2) suggests further that once it is lost, it is not readily regained or reselected. Second, the stability of the precise sequence of the 10-bp site over very long intervals is likewise suggestive of the preservation of an inherited site. Examples include the GCCAGCGGCCC sites associated with ASH and ATH/A TOH genes from fish to mammals (Figs. 1B and C), representing perhaps 450 My of divergence (Peterson et al., 2004); and the constant GCCACCGGAC site in all Dipteran ase genes (Fig. 1D), representing ~235 My (Peterson et al., 2004). Third, in certain instances, the exact conservation of the 10-bp element along with flanking sequences strongly suggests that the sites in question are indeed orthologous. This phenomenon is frequent (Figs. 1B–F), and perhaps to be expected, over shorter evolutionary distances (<100 My), but chordate Ath genes offer an extreme example (Fig. 1H). Here, the repressor binding site and flanking sequences show extraordinary stability over ~450 My; there is little room for doubt that these sequences are orthologous and have been subject to exceptionally strong selection.

As described in the Introduction, the in vivo function of a high-affinity bHLH repressor site was first demonstrated in the case of the Drosophila ase gene, where the motif GCCAGCGGAC confers direct negative regulation by the hairy protein (Ohnaka et al., 1994; Van Doren et al., 1994). The results of other studies suggest that the sequence elements considered here mediate negative transcriptional regulation in deuterostomes as well. (1) Co-transfection assays have defined a 48-bp region immediately upstream of the human ASCL1 (aka hASH1) gene that mediates repression by the HES-1 bHLH repressor in cultured cells (Chen et al., 1997). (2) An in vivo study using lacZ reporter transgenes failed to observe elevated or ectopic expression driven by the mouse Ascl1 (aka Mash1) promoter when the repressor site was mutated along with another element (Meredith and Johnson, 2000). However, the wild-type construct was expressed very weakly, so the sensitivity of this assay is unclear. (3) A recent report revealed a role for Hes-1 not only in direct transcriptional repression of Ascl1, but also in its de-repression and activation (Ju et al., 2004); however, the function of the conserved binding site we have identified was not investigated in this study. The extraordinary evolutionary conservation of this site certainly suggests that the question of its role in mammals should be revisited experimentally.

Conserved cis-regulatory elements and the identification of ancient and abstract developmental functions

Changes in transcriptional regulatory connections are perhaps the major engine for the generation of developmental novelty during evolution (Carroll et al., 2005;
Davidson, 2001; Wray, 2003; Wray et al., 2003). The establishment of new cis-regulatory linkages allows the co-option of genes for new functions; e.g., the deployment of transcription factors (and their existing targets) in new territories and/or the acquisition of new target genes (Davidson, 2001). But even when a particular gene expression specificity is retained in evolution, stabilizing selection on the output of the associated cis-regulatory module may, nevertheless permit rapid changes in transcription factor binding site sequence and organization (Ludwig et al., 2000; Romano and Wray, 2003). Indeed, many recent studies have emphasized the evolutionary fluidity of factor binding sites within developmental regulatory modules (Costas et al., 2003; Dermietzelis and Clark, 2002; Dermietzelis et al., 2003; McGregor et al., 2001; Ruvinsky and Ruvkun, 2003; Stone and Wray, 2001).

The ancient bHLH repressor/proneural gene connection described here is clearly an exception to this paradigm. We suggest that it represents a distinctive class of transcriptional regulatory linkage that may be expected to show remarkable evolutionary stability. Changes in gene regulatory networks can create evolutionary novelty, allowing metazoan clades to diverge from each other developmentally. But some genes, and their regulators, carry out ancient and abstract functions of fundamental utility to all metazoans. (By abstract, or generic, we mean independent of any specific phylogenetic or developmental context.) Proneural genes are in this category; they are themselves ancient and are the key regulators of neural cell fate specification and differentiation in the ectoderm of probably all species with a nervous system, including cnidarians. Moreover, the capacity to pattern proneural gene activity by transcriptional repression, either globally or locally, is an equally generic function, and the bHLH repressors are equally ancient. The preservation of the Hes repressor/proneural gene linkage reflects, we suggest, the fitness advantage conferred by generic capabilities of broad utility to the development of most or all metazoan nervous systems, as diverse as they have become.

A second characteristic of certain transcription factor/target gene relationships may also contribute to their long-term preservation during evolution. If the transcription factor is uniquely positioned in a gene network to subject the target gene to a specific (and again, abstract) mode of regulation, it may be difficult evolutionarily to replace this linkage with another factor. In the present case, it is noteworthy that in species from cnidarians to mammals, multiple Hes-type repressor genes are directly regulated by Suppressor of Hairless/CBF1, the transcription factor downstream of the Notch cell-cell signaling pathway (Bailey and Posakony, 1995; Jarnail et al., 1995). These repressors are thus perhaps uniquely positioned to effect Notch signal-dependent repression of proneural genes, as in the abstract process of “lateral inhibition” (Lai, 2004). This may have helped to insure the durability of the Hes repressor/proneural gene linkage.

It is less apparent why the specific sequences of the bHLH repressor sites we have studied change so slowly with evolutionary time and why only the highest-affinity sites are selected. It may be that high affinity assures high occupancy, and thus robust and reliable repression, and that only a single such site is necessary. The conservation of precise site sequence may reflect in part the constraints imposed if multiple bHLH repressors of slightly different binding specificity must interact functionally with the site. Another, very interesting possibility is that the identity of the repressor(s) recruited by the Hes repressor protein is determined by the binding site sequence (Bianchi-Frias et al., 2004; Leung et al., 2004).

It seems certain that other developmental regulatory linkages of an age similar to that described here remain to be discovered. The systematic identification of such conserved connections will be a powerful means of uncovering the primordial functions of transcription factors and signaling systems, which in turn will illuminate the ancient regulatory and developmental capacities that comprise the minimal, essential metazoan “toolkit.”

Acknowledgments

We are grateful to Par Towb for providing the Nv Hes-1 cDNA and to Stefano Stifani for the rat GST-Hes1 fusion construct. We thank John Huesenbeck for instruction on phylogenetic inference and for helpful discussions. Bill McGinnis and Boris Igic provided valuable critical comments on the manuscript. Supported by NIH grant GM046993 to J.W.P.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ydbio.2005.03.004.

References


Romano L.A., Wray G.A., 2003. Conservation of Endo16 expression in sea urchins despite evolutionary divergence in both cis and trans-


Supplementary Figure. S1. Replicate of phylogenetic tree shown in Fig. 2, with branch lengths and support values included. This tree represents the second of two independent runs of the Bayesian likelihood analysis program (MrBayes 3); except for the placement of one gene, these yielded identical results. In this tree, Am ase branches with Am ASH, whereas it branches with other ase genes in Fig. 2. Numbers at the interior nodes are the estimated
[Supplementary Figure S1 continued] posterior probabilities that the clades shown are valid. Red star indicates the node that roots the tree in Fig. 2, separating the achaete-scute (ASH) subfamily from the atonal (ATH/ATOH) subfamily. Species names are listed in legends for Figs. 1 and 2.
Chapter Five, in full, is a reprint of the following publication. Rebeiz M, Stone T, Posakony JW. An ancient transcriptional regulatory linkage. *Dev Biol.* 2005 May 15;281(2):299-308. I am the primary researcher and author and James W. Posakony directed and supervised the research that forms the basis for this chapter. Tammie Stone performed the electrophoretic mobility shift assays.