Bioinformatic Analysis of Neurofibromatosis Type 1 on Transcriptional Regulation

by

TSZ KIN (BERNARD) LEE

B.Sc., Simon Fraser University, 2001

A THESIS SUBMITTED IN PARTIAL FULFILMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE STUDIES

(Faculty of Medicine; Department of Medical Genetics)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

August 2003

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Medical Genetics

The University of British Columbia
Vancouver, Canada

Date August 28, 2003

Abstract

The objective of this study was to identify potential transcription factor binding sites in the human NF1 gene through phylogenetic footprinting. The 5' upstream region (5UR) and Exon 1-Intron 1 of the NF1 gene from human, mouse, rat, and pufferfish were compared and analyzed using various bioinformatic tools. Three regions that have equal or higher homology than the coding regions were discovered in the NF1 5UR, and four more very highly homologous regions were found in intron 1. Five of these highly homologous regions had transcription factor binding site predictions that were similar for the two binding site detection programs used in this study. One of the highly homologous regions within intron 1 had no shared predictions between the two transcription binding site detection programs. Another highly homologous region in intron 1 is a tetranucleotide repeated sequence. One of the highly homologous regions in the 5UR that contains several transcription factor binding site predictions spans the transcription start site. This region includes a 24bp sequence acttccggtggggtgtcatggcgg 310-333 bp upstream of the translation start that is identical in human, mouse and rat and differs by only1 bp in *Fugu*. This sequence may contain the core promoter responsible for NF1 transcription initiation.

TABLE OF CONTENTS

LIST OF FIGURES

(electronic version available in Appendix II)

LIST OF TABLES

(electronic version available in Appendix II)

ABBREVIATIONS

5' UTR	5' Untranslated Region
5UR	5' Upstream Region
ACTA1	Alpha-Skeletal Actin 1
AP1	Activation Protein
BLAST	Basic Local Alignment Search Tool
BLAT	BLAST-Like Alignment Tool
BRE	TFIIB Recognition Element
BTEB1	Basic transcription element B1
CAL	Café-au-lait spot
CRE	cAMP Resonse Element
CREB	Creb-binding protein
DPE	Downstream Promoter Element
DPF	Dragon Promoter Finder
EI1	Exon 1 Intron 1
EKLF	Erythroid kruppel like factor
EPD	Eukaryotic Promoter Database
GAP	GTPase Activarting Protein
GFF	Gene-Finding Format
GTF	General Transcription Initiation Factor
HBB	beta-globin
HHR	Highly Homologous Region
Inr	Initiator
LCK	Lymphocyte-specific protein-tyrosine kinase
LOH	Loss Of Heterozygosity
MPD	Myeloproliferative disease
MPNST	Malignant Peripheral Nerve Sheath Tumour
MT1	Metal Response Element
mVISTA	main Visualization Tool for Alignment
MyoD	Myoblast determination gene product
NF1	Neurofibromatosis 1
NF1GRD	NF1 GAP-Related Domain
NF1HCS	NF1 Highly Conserved Sequence
ORF	Open Reading Frame
Pax	Paired Box
PET	Positron Emission Tomography
PI3K	Phophoninositol-3'-kinase
PIC	a pre-initiation complex (
Rfam	RNA family database
SnRNA	Small nuclear RNA
Sp1	Simian-virus-40-Protein-1
SRE	Serum Response Element
TAF	TBP-associated factors
TAF7	TBP-associated factor 7
TBP	TATA-binding protein
TFIID	Transcription factor D
TFAP2C	Transcription factor AP-2 gamma
TLP	TBP-like proteins

TNFα	Tumour necrosis factor α
TPA	phorbol ester 12-*O*-tetradecanoylphorbol-13-actate response element
TRE	TPA-response element
TRF	TBP-related factors
TRRD	Transcription Regulatory Regions Database
TSG	Tumour Suppressor Gene
TSS	Transcription Start Site
vMyb	oncogene of Avian myeloblastosis virus

ACKNOWLEDGEMENTS

I would like to thank and express my gratitude:

to my supervisor Dr. J. Friedmen for his advice and guidance throught this project,

to all the members of the Friedman laboratory for making my M.Sc. fun,

to Drs. S. Jones, P. Hieter, and A. Rose for serving on my graduate committee,

to G. Robertson and the Sockeye group of the BC Genome Sequence Centre for developing such a wonderful program,

to Dr. W. Wasserman and Harry Joe for their valuable advice,

to my parents and family for their continuous support.

Chapter 1. Introduction

1.1 History of Neurofibromatosis 1

One of the earliest records of neurofibromatosis 1 (NF1) may be the stone renderings identified from 300 BC that appear to describe neurofibromas, one of the many manifestations of NF1 (Zanca *et al.*, 1980). Throughout the centuries, there were other descriptions of the disease, but most concentrated on dermological features, with occasional attention to the familial aspects of NF1. It was not until 1882 that Frederick von Recklinghausen gave a full description on NF1 phenotype. He also coined the term "neurofibroma" for NF1 tumours observed in the skin, recognizing their origin from fibrous tissues surrounding small nerves (Von Recklinghausen, 1982).

In 1956, the landmark study of Crowe, Schull, and Neel (Crowe *et al.*, 1956) described the high incidence and high spontaneous mutation rate, as well as the usefulness of the café-au-lait spot as a diagnostic feature and the wide range of clinical features that can occur (Lynch *et al.*, 2002). In 1987, the National Institutes of Health organized a Consensus Panel that established a set of diagnostic criteria for NF1, which led to reliable and consistent diagnoses for affected individuals. With identification of the *NF1* gene (Viskochil *et al.*, 1990; Wallace *et al.*, 1990) and its protein, neurofibromin (Gutmann *et al.*, 1991), physicians and researchers began to understand the pathogenesis of NF1 better, but effective therapies for individuals who suffer from the disease remain elusive.

1.2 Neurofibromatosis 1: Clinical Features

NF1 is an autosomal dominant disease with 100% penetrance (Littler, 1990; Viskochil, 2002). The prevalence is 2 to 3 per 10,000, and it can affect individuals of any age, gender, race, or ethnic background (Friedman, 2002). The two main characteristics of NF1 are its progressive nature and its variability. Different patients, even ones from the same family, can have a wide range of disease manifestations with different levels of severity. Some of the key NF1 manifestations are: café-au-lait spots, axillary and other intertriginous freckling, Lisch nodules, neurofibromas, malignant peripheral nerve sheath tumours, optic pathway gliomas, and skeletal dysplasia.

1.2.1 Café-au-lait spots

Café-au-lait spots (CALs) occur in almost all children with NF1 in infancy. The size and number of CALs usually increase in the first two years of life (Figure 1). These macules can reach 10 to 30 mm in diameter in adult NF1 patients. They are usually ovoid in shape with uniform pigmentation that varies in intensity based on background cutaneous pigmentation (Friedman, 2002). Café-au-lait spots can occur anywhere on the body of NF1 patients except the scalp, eye-brows, palms, and soles, and range in number from six to several dozens. Histologically, an increased number of macromelanosomes is found within melanocytes of café-au-lait spots of NF1 patients, but this is not diagnostic for NF1 (Konrad, *et al.*, 1974; Bhawan *et al.*, 1976; Martuza *et al.*, 1985).

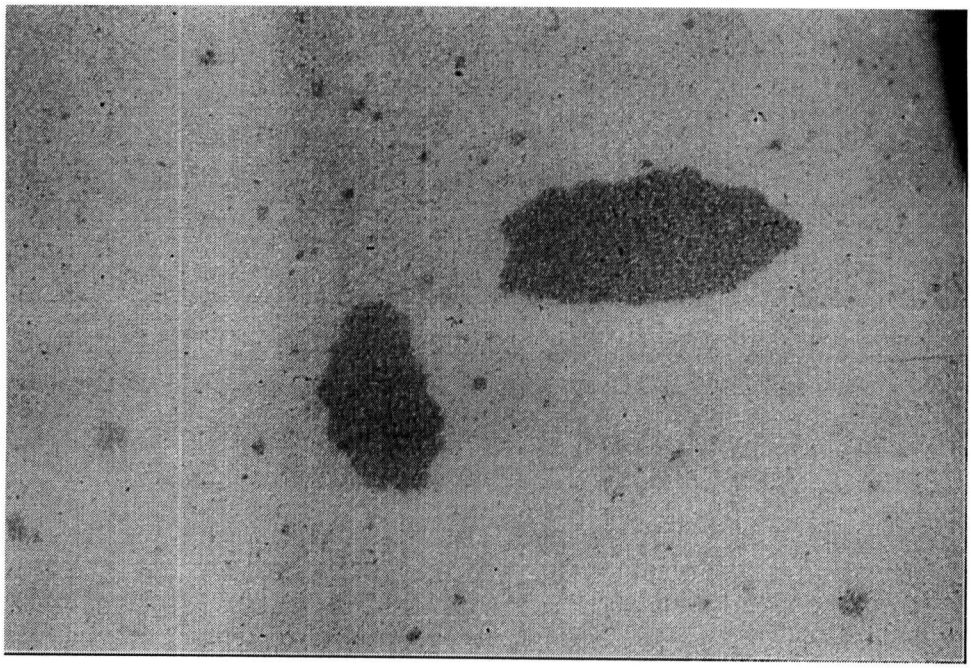


Figure 1. Café-au-Lait Spots.

1.2.2 Axillary and Other Intertriginous Freckling

Crowe was the first to point out the appearance of freckles in the axillary and inguinal regions (Figure 2) as a useful diagnostic feature (Crowe, 1964). Although these freckles are similar to café-au-lait spots in colour, they are smaller and generally appear in clusters. Other common sites of freckling in NF1 include the upper eyelids, face, trunk, and proximal extremities.

1.2.3 Lisch Nodules

In 1937, a Viennese ophthalmologist named Karl Lisch noted that raised, pigmented iris hamartomas (Figure 3) were a frequent clinical manifestation in NF1 patients (Riccardi *et al.*, 1986). Lisch nodules are more common in NF1 patients older than 10 years (Otsuka *et al.*, 2001). Lisch nodules are not true tumours and have no effect on vision. They are generally not associated with any clinical symptoms. Because most adults with NF1 have Lisch nodules and they are highly characteristic, Lisch nodules are useful as a diagnostic feature.

1.2.4 Neurofibromas

Neurofibromas are benign tumours composed of Schwann cells, fibroblasts, perineurial cells, axons, and mast cells embedded in extracellular matrix (Ferner, O'Doherty , 2002). Discrete dermal and plexiform neurofibromas are the two main forms.

Clinically, dermal neurofibromas are present in most adult NF1 patients as discrete masses arising from a single nerve as cutaneous or subcutaneous tumors (Figure 4). Although they can cause pain, discomfort, itching, and/or emotional concern because of their cosmetic effects, dermal neurofibromas are rarely if ever associated with other neurological symptoms or malignant change. Their growth pattern is variable and unpredictable, but they tend to increase in both size and number as an NF1 patient ages. Furthermore, since the onset of dermal

Figure 2. Axillary Freckling. (Wainer S. A child with axillary freckling and cafe au lait spots. CMAJ. 2002 Aug 6;167(3):282-3.)

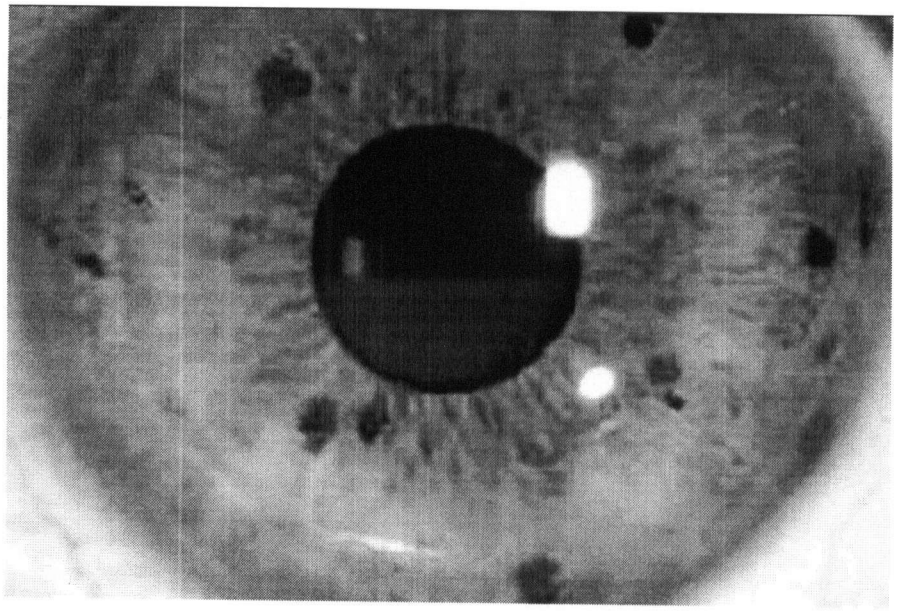


Figure 3. Lisch Nodules.

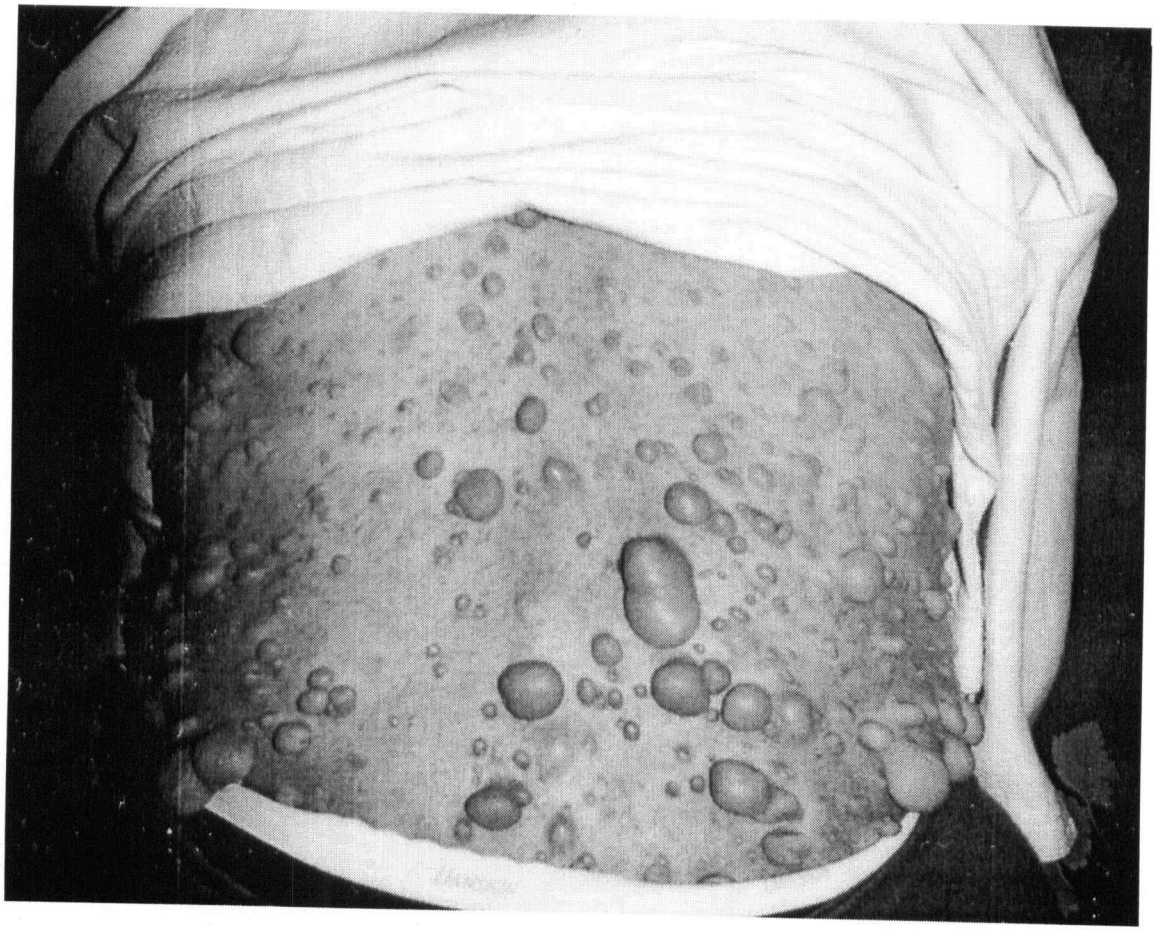


Figure 4. Discrete Cutaneous Neurofibromas

neurofibromas is usually just before puberty and an increased growth rate is observed during pregnancy, neurofibromas may be under hormonal influence (Dugoff *et al.*, 1996).

Unlike dermal neurofibromas, diffuse plexiform neurofibromas are congenital lesions (Figure 5). Diffuse neurofibromas may involve major and minor nerves, muscle, connective tissue, and overlying skin (Wiestler *et al.*, 1994). Trunk, limbs, head, and neck are all common locations for diffuse plexiform neurofibromas. Furthermore, because congenital diffuse plexiforms often extend fingerlike projections into surrounding tissues, surgical removal is close to impossible without sacrificing adjacent normal tissues. On the other hand, nodular plexiform neurofibromas are confined within the perineurium; they may involve major or minor nerves. The sizes vary greatly, and they can extend the entire length of a nerve (Friedman *et al.*, 1999). They are usually asymptomatic and can go unobserved for many years.

1.2.5 Malignant Peripheral Nerve Sheath Tumours (MPNSTs)

The overall lifetime rate for NF1 patients to develop MPNSTs (Figure 6) is about 10% (Evans, *et al.*, 2002). Patients with extensive and centrally-located plexiform neurofibromas, previous history of radiotherapy, a family history of cancer, or microdelection of the *NF1* locus may warrant closer monitoring by physicians for the development of MPNSTs (Ferner, Gutmann, 2002). MPNSTs are highly aggressive and are often fatal. Earlier detection may be possible with 18-fluoro-deoxyglucose-positron emission tomography (PET), which may distinguish benign and peripheral nerve sheath tumours (Ferner *et al.*, 2000). Surgical removal is the only effective way to treat MPNSTs that is currently available (Ferner, Gutmann, 2002; Thomas *et al.*,

Figure 5. Diffuse Plexiform Neurofibroma of the right leg.

Figure 6. Local recurrence of MPNST in a 41 year-old female with NF1 4 months after radical resection (Stark *et al.*, 2001)

1983), although chemotherapy using doxorubicin and/or ifosfamide may offer palliation and sometimes long-term remission (Santoro *et al.*, 1995).

1.2.6 Optic Pathway Gliomas

Optic pathway gliomas, or optic gliomas, usually appear in NF1 patients during the first 5 years of their lives (Figure 7). These tumours usually involve the intraorbital portion of both optic nerves, and at least one-half are asymptomatic in NF1 patients. Symptomatic optic gliomas often manifest as visual abnormalities with decreased visual acuity, poor colour vision, optic atrophy, or abnormal pupillary function (Listernick *et al.*, 1995). Therapies are generally considered only when there is a progressive loss of vision or progressive proptosis (forward displacement or projection of the eyeballs). Because of potential neurological damage and endocrinological disturbance, radiotherapy is not recommended. Chemotherapy may be the best choice in controlling progressive optic gliomas in NF1.

1.2.7 Skeletal Dysplasia

Although about 10 percent of patients with NF1 develop scoliosis (Figure 8), dysplastic scoliosis is fairly uncommon. Dysplastic scoliosis is characterized by a sharp curve over a few vertebrae and is a serious and characteristic clinical manifestation of NF1. Neurologic complications may result from spinal cord compression but can often be prevented or alleviated by surgery (Friedman, 2002; Korf, 2002).

Congenital tibial dysplasia in NF1 patients leads to thinning of the cortex of the bone and anterolateral bowing of one leg (Figure 9). Since the bone is bowed and weakened, it is

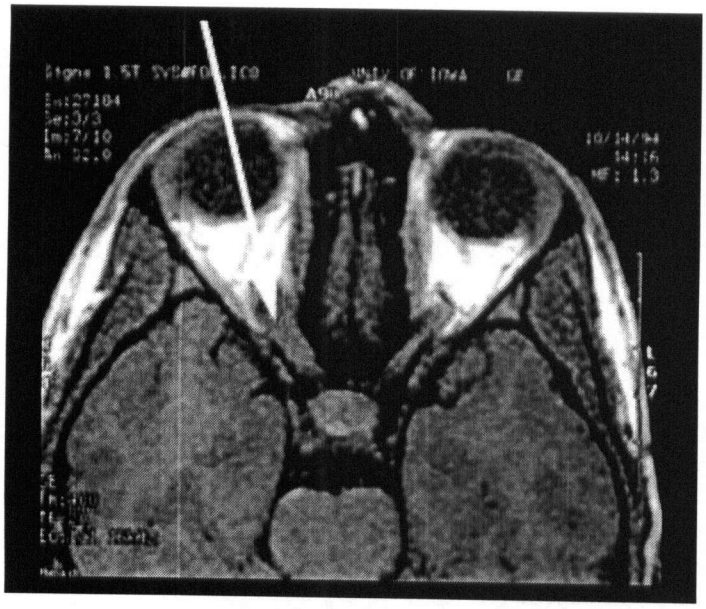


Figure 7. Optic Glioma in a patient with NF1

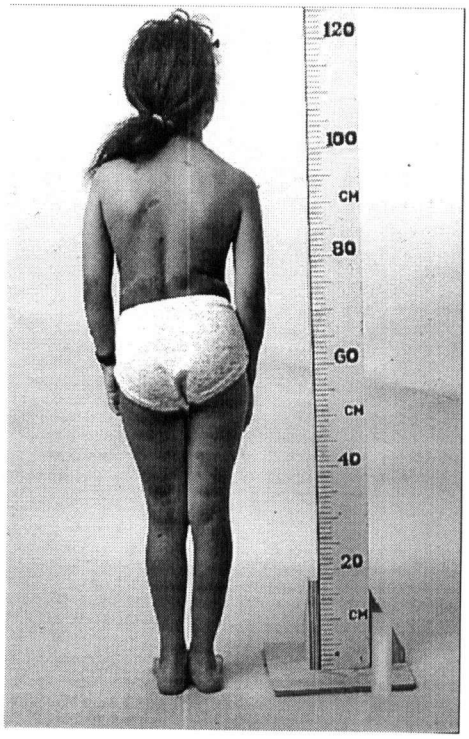


Figure 8. Scoliosis in a girl with NF1.

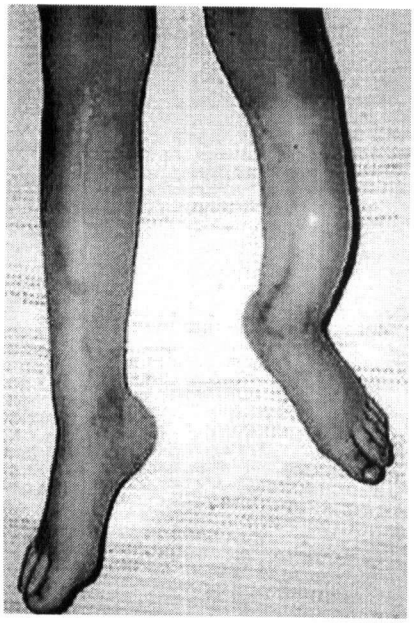


Figure 9. Tibial dysplasia in a child with NF1

(http://cc.oulu.fi/~anatwww/NF/Neurofibromatosis/)

vulnerable to fracture, followed by poor healing and often pseudarthrosis, which is the inability to form normal callus for healing (Friedman, 2002; Korf, 2002).

Another type of dysplasia that may occur in NF1 is sphenoid wing dysplasia (Figure 10). It is unilateral and sometimes is associated with an orbital plexiform neurofibroma. Sphenoid wing dysplasia usually has little clinical consequence, but it may progress and compromise the integrity of the bony orbit. In some serious cases, pulsating enopthalmos, which is a sucken eyeball or herniation of the brain into the orbit can occur (Friedman, 2002; Korf, 2002).

1.3 National Institutes of Health Neurofibromatosis 1 Diagnostic Criteria

As mentioned, NF1 was well described as a clinical entity by Frederick von Recklinghausen in 1882. Standardized diagnostic criteria for NF1 were established by a Consensus Panel organized by National Institutes of Health in 1987 to facilitate linkage studies (Table 1; National Institutes of Health Consensus Development Conference, 1988). In 1997, the NIH Diagnostic Criteria were re-evaluated, and continued use without modification was recommended for both clinical diagnosis and research (Gutmann *et al.*, 1997). Although the NIH Criteria have high sensitivity and specficity in adults, these criteria cannot always be used with confidence on young children, because many young children who do not meet the NF1 criteria later develop unequivocal NF1 (DeBella *et al.*, 2000). This problem is especially apparent in children with sporadic NF1 because, with the exception of CALs, which are present in almost all NF1 patients during infancy, most NF1 features are uncommon in children (Friedman *et al.*, 1997). In contrast, children with familial NF1 only require CALs to meet the NIH diagnostic criteria because these children have an affected first-degree relative.

Figure 10. Sphenoid Wing Dysplasia compromising the bony orbit in a patient with NF1 (http://www.neurorad.ucsf.edu/previouscases/03012002.html).

Table 1. National Institutes of Health Diagnostic Criteria for Neurofibromatosis 1.

Neurofibromatosis 1 is present in a patient who has two or more of the following signs:

- Six or more café-au-lait macules more than 5mm in greatest diameter in prepubertal individuals or more than 15mm in greatest diameter after puberty
- Two or more neurofibromas of any type or one or more plexiform neurofibroma
- Freckling in the axillary or inguinal regions (Crowe's sign)
- An optic pathway tumor
- Two or more Lisch nodules (iris hamartomas)
- A distinctive osseous lesion, such as sphenoid wing dysplasia or thinning of the cortex of the long bones (with or without pseudarthrosis)
- A first-degree relative (parent, sibling, or offspring) with neurofibromatosis 1 by the above criteria

From NIH Consensus Development Conference, Neurofibromatosis: Conference Statement (National Institutes of Health Consensus Development Conference, 1988)

1.4 Neurofibromatosis 1 and Associated Genetic Principles

NF1 is an autosomal dominant Mendelian disease affecting 2 to 3 people per 10,000 worldwide, regardless of gender, race, or ethnic background (Lakkis *et al.,* 2000; Viskochil, 2002). Affected individuals are heterozygous for an *NF1* mutation. Constitutional homozygosity for an *NF1* gene mutation has never been reported in humans. Mice homozygous for a targeted mutation in *NF1* die *in utero* between embryonic days 12.5 and 13.5 because of cardiac defects (Dasgupta *et al.,* 2003; Lakkis *et al.,* , 1999). Homozygous mutation of the *NF1* locus in humans is probably lethal, as it is in mice (Friedman, 1999). Among NF1 patients, 30 to 50 percent of cases are sporadic with no family history of NF1. These data translate into an extremely high mutation rate of approximately 10^{-4} per generation. Most *de novo* mutations in the *NF1* gene involve the paternal chromosome (Jadayel *et al.,* 1990), although whole-gene microdeletions tend to involve the maternal chromosome (Upadhyaya *et al.,* 1998).

1.4.1 Penetrance

NF1 as a disease is 100 percent penetrant (Littler *et al.,* 1990), but its various manifestations are incompletely penetrant and often age dependent (Viskochil, 2002). For example, CALs are usually present in infants with NF1, while axillary freckling, Lisch nodules, and discrete dermal neurofibromas usually do not appear until a patient gets older.

1.4.2 Variable Expressivity

NF1 exhibits highly variable expressivity. A wide spectrum of severity and features is common within a NF1 family, and parents, siblings, and offspring can have completely different disease manifestation and severity (Friedman, 2002).

On the other hand, intrafamilial variation is smaller than interfamilial variation. For example, one study of monozygotic twins found the highest correlation in the number of CALs and neurofibromas in monozygotic twins, followed by first-degree relatives, and the lowest correlation in more distant relatives (Easton *et al.,* 1993). Another study found that certain NF1 manifestation are more likely to be shared between first-degree relatives, between siblings, or between parents and children (Szudek *et al.,* 2002). For example, Lisch nodules and CALs were more strongly associated between first-degree relatives than between second-degree relatives. Therefore, besides the specific mutation in *NF1* gene, other factors and mechanisms (*e.g.* epigenetic factors) that can control or affect NF1 manifestations may be shared within a family.

Epigenetic factors are modifications of DNA that can alter gene expression without altering the nucleotide sequences of that gene. For example, imprinting and gene silencing defects can lead to Prader-Willi and Angelman syndromes (Vogels *et al.,* 2002). Although methylation of the *NF1* gene has been observed during different stages of development (Haines *et al.,* 2001), there is no evidence that methylation can cause NF1.

1.4.3 Pleiotropy

Since CALs, axillary freckling, and Lisch nodules all involve cells that are of neural crest origin (Benish, 1975) and there is a high level of NF1 gene expression in embryonic neural crest tissue, NF1 is sometimes considered to be a neurocristopathy (Stocker *et al.,* 1995), i.e. a developmental anomaly of neural crest-derived tissues. However, NF1 can also manifest as bony dysplasia, vasculopathy, and cognitive abnormality. So, NF1 can involve tissues of ectodermal, endodermal, or mesodermal origin, rather than just neural crest tissue. The *NF1* gene is ubiquitously expressed (Gutmann *et al.,* 1995), so it is not surprising that NF1 is a pleiotropic disorder.

1.4.4 Mosaicism

Mosaicism is the occurrence of two or more cell populations of different constitutions that are all derived from a single zygote. Mosaicism for a *NF1* mutation can cause "segmental NF1", which occurs when NF1 features are limited to a localized body region (Tinschert *et al.,* 2000). Although individuals with segmental NF1 have a lower risk of developing medical complications compared to the common form of NF1, they are at a higher risk than the general population for having a child with NF1 because of the possibility of germline mosaicism. There has also been a report of a clinically normal individual having germline mosaicism that caused NF1 in two of his children (Lazaro *et al.,* 1994).

1.5 Neurofibromatosis 1 Genetics

1.5.1 Basic Gene Structure

In 1987, the NF1 gene was found to be closely linked to a marker pHHH202 (D17S33), which was later located at chromosome 17q11.2-12 by physical mapping (White *et al.,* 1987; van Tuinen *et al.,* 1987). This was also supported by the reports of two balanced chromosomal rearrangements t(1;17)(p34.3;q11.2) and t(17;22)(q11.2;q11.2) in NF1 patients (Schmidt *et al.*, 1987; Ledbetter *et al.,* 1989). In 1990, the *NF1* gene was identified through the detection of deletion mutations from NF1 patients and human-mouse homology (Viskochil et al., 1990), the identification of splice junctions and sequencing of exons (Cawthon *et al.,* 1990), and the identification of the ubitquitously-expressed *NF1LT* (*NF1*) transcript (Wallace *et al.,* 1990). According to ENSEMBL version 12.31.1, the human *NF1* gene contains 58 exons and spans 279317 bp, with an open reading frame of 8520 bp and a protein (neurofibromin) size of 2839 amino acids. However, there are three in-frame alternatively-spliced variants, exon 23a (63bp),

exon 9a (30 bp) and exon 48a (54 bp) (Viskochil *et al.*, 2002), and ENSEMBL only includes the first of these. In other words, the *NF1* gene has a total of 60 exons. The features are summarized in Figure 11. The NF1 transcription start site is considered to be 484 bp upstream of the translation start site, which is the beginning of the ORF (Marchuk *et al.*, 1991; Hajra *et al.* 1994). The 3' UTR has a length of 3.5 kb (Li *et al.*, 1995). The NF1 promoter is thought to lie within a CpG island that is 471 bp long (Rodenhiser *et al.*, 1993) and starts at 731 bp upstream of the translation start site according to UCSC. The NF1 promoter does not include a TATA box or CCAAT box (Viskochil, 1999).

1.5.2 Neurofibromin and the Ras Pathway

NF1 patients are predisposed to develop neurofibromas, MPNSTs, and leukemia. Because NF1 is an autosomal dominant disease and because loss of heterozygosity (LOH) in *NF1* gene activity has been observed in tumour cell lines (Legius *et al.*, 1993; Sawada *et al.*, 1996; Side *et al.*, 1997), it was hypothesized that the *NF1* gene product, neurofibromin, may function as a tumour suppressor gene in certain tissues. Furthermore, about 360 amino acids of neurofibromin, coded by exons 21 to 27a, share homology with a Ras-specific GTPase activating protein (GAP), p120GAP, and the yeast homologues IRA1 and IRA2 (Xu, O'Connell *et al.*, 1990; Trahey *et al.*, 1987; Tanaka *et al.*, 1990). This segment, which is now called NF1 GAP-related domain (NF1GRD), has been shown to stimulate GTP-hydrolysis of normal Ras but not oncogenic Ras in yeast (Xu, Lin *et al.*, 1990; Scheffzek *et al.*, 1998).

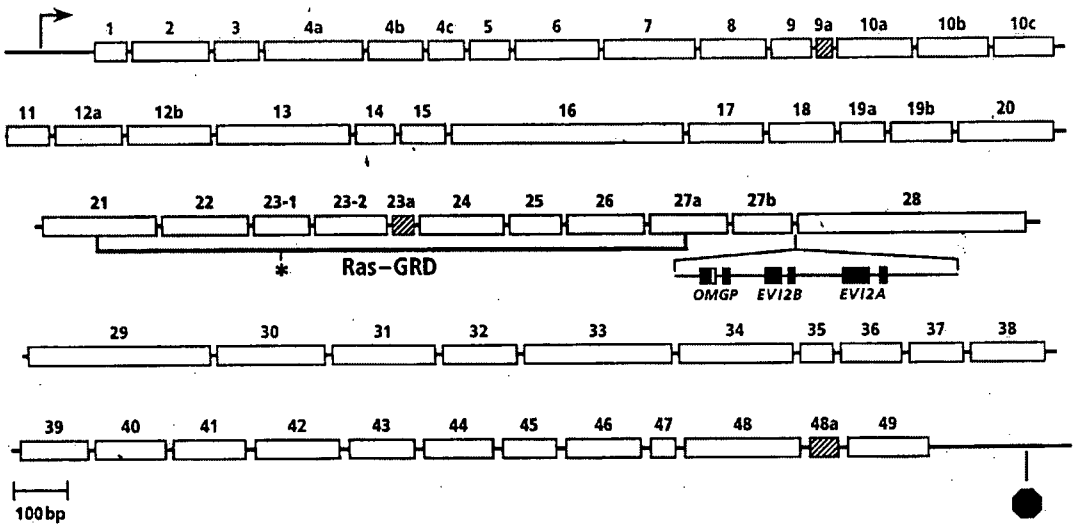


Figure 11. Schematic drawing of *NF1* exons and embedded genes. Intron are not shown to scale. The scale in the lower left corner is for the size of exons. The transcription start site is depicted as a horizontal arrow upstream of exon 1. The transcription stop site and polyadenlyation site are marked with an octagon. The GAP-related domain, *ras*-GRD, is shown spanning exon 21 to 27a (Scheffzek *et al.*, 1998). The alternative splice forms are in-frame insertions of exon 9a, 23a, and 48a, and they are hatched. The embedded genes are shown in bold in intron 27b and are transcribed in the opposite direction (telomere-to-centromere). The t(1:17) and t(17:22) translocation breakpoints lie in intron 27b, upstream of *OMGP*, and in intron 31, respectively. The asterisk in exon 23-1 represents a site of mRNA processing, C3916U, that leads to premature truncation at codon 1303 (Cappione *et al.*, 1997). Picture obtained from Viskochil, 1999.

Being central to cellular growth and differentiation, Ras protein activity is under tight regulation and cycles between active GTP-bound conformation (Ras-GTP) and inactive GDP-bound conformation (Ras-GDP). Ras-GTP can stimulate cell proliferation by activating MEK (formerly called MAP kinase kinase) and inhibit apoptosis by activating phophoinositol-3'-kinase (Wittinghofer, 1998). On the other hand, GAPs act as negative control for Ras activity by enhancing the slow intrinsic GTPase activity of Ras and increasing GTP hydrolysis rate (Figure 12). Since neurofibromin contains the NF1GRD, it was hypothesized to have a regulatory and tumour suppressing role in the Ras-MAP kinase pathway (Basu *et al.,* 1992).

Neurofibromin's regulatory role on tumour growth is supported by the observed abundance of active GTP-bound Ras and the absence of functional neurofibromin in malignant tumours of NF1 patients (DeClue *et al.,* 1992). Many other studies have also shown the relationship between neurofibromin and the RAS pathway. For example, a study using an *in vivo* RAS-binding fluorescence assay has shown that loss of neurofibromin is associated with an increase in RAS activity in neurofibroma Schwann cells (Sherman *et al.,* 2000). Studies on MPNSTs have shown that hyperactivation of RAS can be achieved with the loss of neurofibromin and mutations of p53, $p16^{INK4a}$, and $p14^{ARF}$ tumour suppressor genes as well as $p27^{Kip1}$ cell-cycle growth regulator (Kourea, Orlow *et al.,* 1999; Kouea, Cordon-Cardo *et al.,* 1999). NF1-related myeloid leukemias and pilocytic astrocytomas have also been demonstrated to have loss of neurofibromin expression and increased RAS pathway activation (Side *et al.,* 1997; Gutmann *et al.,* 2000; Lau *et al.,* 2000).

Mouse models of NF1 also agree with findings from cell culture. For example, although $NF1^{+/-}$ mice do not develop neurofibromas or astrocytomas, these mice develop leukemia and

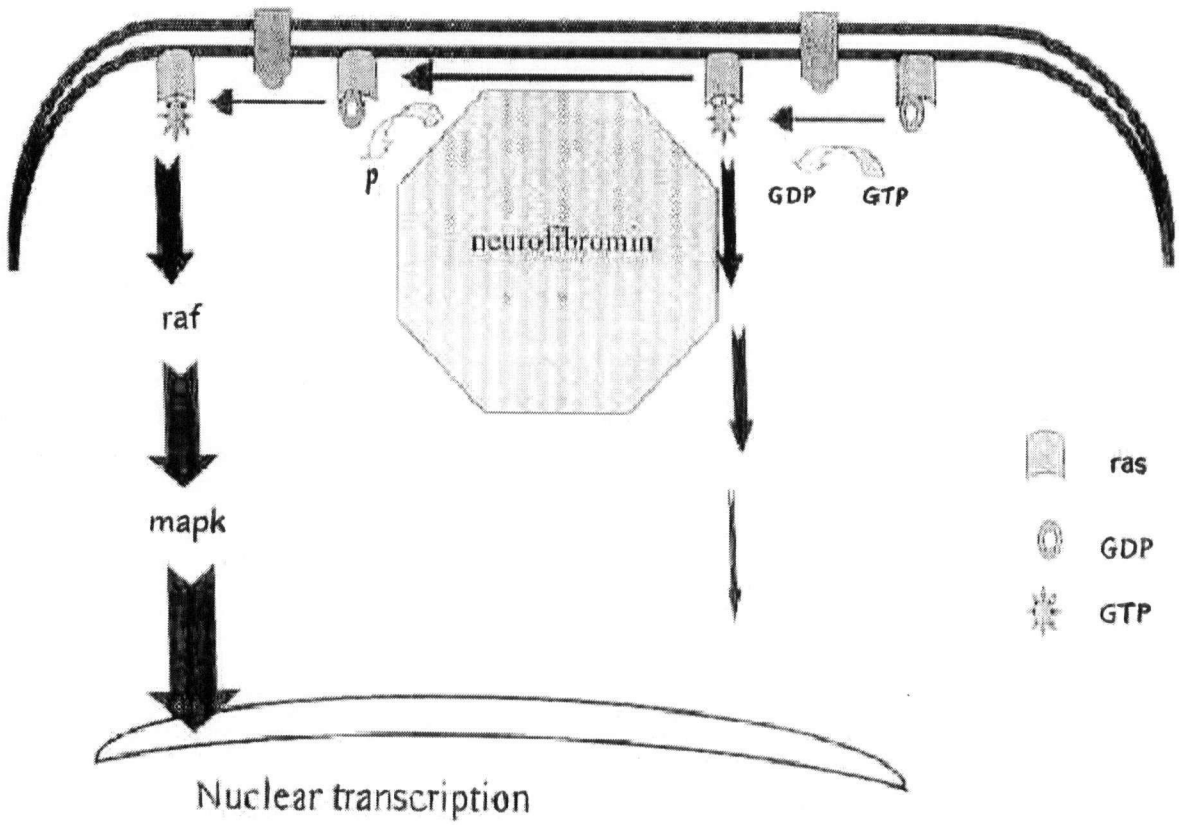


Figure 12. Neurofibromin acts as a negative regulator of ras signal transduction. GDP = guanosine diphosphate; GTP = guanosine triphosphate (Viskochil D. Genetics of neurofibromatosis 1 and the NF1 gene. J Child Neurol. 2002 Aug;17(8):562-70; discussion 571-2, 646-51.)

myeloproliferative disease (MPD). Through adoptive transfer of *NF1*$^{-/-}$ fetal liver cells, mice myeloid lineage cells have been shown to be hypersensitive to the proliferative factor GM-CSF resulting in activation of the Ras pathway (Zhang *et al.,* 1998). *NF1*$^{+/-}$ mice also have increased numbers of brain astrocytes with cancer-like characteristics. For example, studies have shown that these cells possess abnormal spreading, attachment, and motility properties, cell-autonomous growth advantage, and increased RAS pathway activation (Gutmann *et al.,* 1999; Gutmann *et al.,* 2001; Bajenaru *et al.,* 2001).

1.6 *NF1* Gene Mutations

As mentioned, NF1 is a very large gene of 279317 bp with an unusually high mutation rate of 10^{-4} per generation. The size of the *NF1* gene allows many possible mutation sites. Furthermore, up to half of NF1 patients represent new mutations. Most of these mutations are novel mutations and have been described affecting amost every exon (Upadhyaya *et al.,* 1998). Some particular exons may be more prone to mutation than others (Messiaen *et al.,* 1999; Ars *et al.,* 2000; Fahsold *et al.,* 2000).

NF1 gene mutations can also affect introns, disrupting splicing patterns. For example, mutations at 5' splice sites of introns 14 and 16 and a mutation at the 3' splice site of intron 31 have been reported (Origone *et al.,* 2003; Maynard *et al.,* 1997; Ainsworth *et al.,* 1994; Hatta *et al.,* 1995). No mutation in the promoter region or 5' upstream region of the *NF1* gene has been reported to date. However, since the promoter region is crucial to transcription initiation and most transcription factors are concentrated upstream of a gene, mutations in these regions are expected to have a negative impact on *NF1* gene transcription and expression.

Furthermore, microdeletion at the *NF1* locus can lead to more severe cognitive abnormalities and higher risk of MPNST development, and it has been suggested that genes adjacent to the *NF1* gene may also interact with the *NF1* gene or its product (Leppig *et al.,* 1997).

1.7 General Transcription and Transcription Factors

Transcription, the synthesis of RNA molecules from DNA, is mediated by RNA polymerases, which can generate mRNA for protein synthesis, ribosomal RNA for ribosome formation, transfer RNA for translation, and other RNA molecules for structural, catalytic, or regulatory functions. There are three types of RNA polymerase in eukaryotic cells - RNA polymerases I, II, and III. RNA polymerase I is confined to the nucleolus and is responsible for the transcription of 18S, 5.8S, and 28S rRNA. RNA polymerase III is responsible for producing 5S rRNA, tRNA molecules, 7SL RNA and some of the snRNA molecules essential for splicing. Lastly, RNA polymerase II transcribes all polypeptide - coding genes and some snRNA genes (Strachan *et al.,* 1999a)

In general, RNA polymerases are free floating, and they may slide along a chromosome. Transcription usually proceeds only when there is activation by various distal or proximal transcription activators. This activation localizes some other coactivators, and together they promote RNA polymerase basal transcription factor complexes, which usually contain TATA-binding protein (TBP) and TBP-associated factors (TAFs), to bind to the core promoter (Pugh, 2000). These complexes finally recruit RNA polymerase to form a tight binding on DNA sequences (Figure 13). This is then followed by DNA unwinding, RNA chain elongation, and eventually termination, which is mediated by a special termination signal in the DNA. This

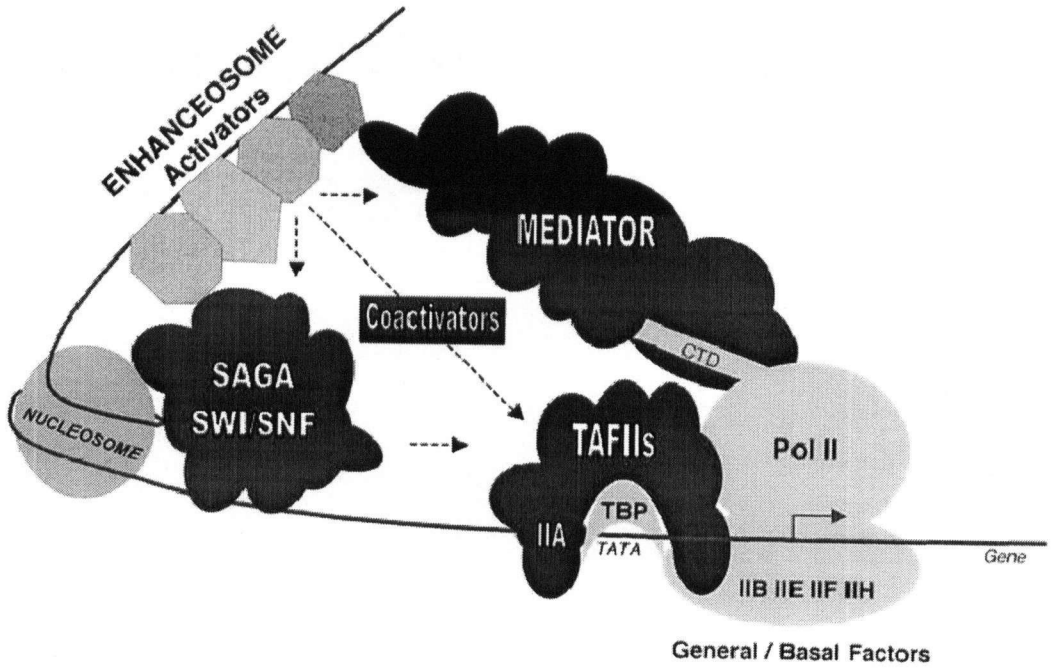


Figure 13. Optimal induction of gene transcription by activators involves various coactivators and protein-protein interactions. Coactivators (dark blue) are recruited by promoter-bound activators (pink, polygons) to remodel chromatin structure (nucleosome, green) and/or to stimulate the recruitment or activity of the general transcription machinery (yellow, general/basal factors) during initiation of transcription by RNA polymerase II (pol II) at the core promoter or during transcription elongation. Coactivators include (1) proteins and complexes that can intimately associate with (or be part of) the general transcription machinery, *e.g.*, TBP-associated factors (TAFIIs) of the TFIID complex, TFIIA (IIA), and the Pol II-associated SRB/Mediator Complex (Mediator) and (2) chromatin-modifying/remodeling factors and complexes that modulate the generally repressive influence of chromatin on protein-DNA interactions (*e.g.*, SAGA histone acetylase and SWI/SNF ATP-dependent nucleosome remodeling complexes). Note that multiprotein cofactor complexes (*e.g.*, the TAFII-containing complexes TFIID and SAGA) might be involved both in chromatin modification (histone acetylation) and in interactions with activators and the general transcription machinery. (Martinez E. Multi-protein complexes in eukaryotic gene transcription. Plant Mol Biol. 2002 Dec;50(6):925-47.)

termination signal is different from the stop codon used for transcription. Lastly, RNA polymerase stops and releases both the DNA and RNA. For RNA polymerase II products except snRNA, the released RNA goes through subsequent modifications like capping, splicing, and poly adenylation and eventually becomes mRNA (Cramer *et al.,* 2001).

Transcription regulation for RNA polymerase II is tissue- and stage- specific through the orchestration of three groups of factors (Martinez, 2002):

1. Sequence-specific DNA-binding transcription factors and regulators, which can be proximal to the promoter or distal.
2. Ubiquitious factors like RNA polymerase II, TATA-binding protein (TBP), TBP-related factors (TRFs), general transcription initiation factors (GTFs) like RNA polymerase II basal transcription factor D (TFIID) complex, and core promoter DNA elements (*e.g.* TATA box, TC-rich promoter, initiator, DPE).
3. Coactivators and corepressors like TBP-associated factors (TAFs) of TFIID, SAGA histone acetylase, and SWI/SNF ATP-dependent nucleosome remodeling complexes.

The focus of this research is on DNA sequences that potentially contain the *NF1* core promoter and transcription factor binding sites.

1.8 Core Promoter and Promoter Elements for RNA Polymerase II

The promoter, sometimes called the promoter region, is the ultimate target of all transcription regulatory control (Figure 13). The core promoter is generally located within 40 bp upstream or downstream (-40 to +40) of the transcription start site, which is designated as +1. There is also a proximal promoter region from -40 to -250 relative to the +1 transcription start site, where

several DNA binding factors may bind and influence transcription (Kadonaga, 2002). Within the core promoter, there is a short DNA sequence called the core promoter element, where an RNA polymerase II basal transcription factor (TFII) binds and then recruits RNA polymerase to begin transcription. There are various different core promoter elements like the TATA box, BRE, initiator element, and DPE, and there are different TFII complexes for each.

TATA-box was the first identified and is the probably the best studied core promoter element (Mathis *et al.,* 1981). It has a consensus sequence of TATAAA and is usually located -25 to -35 relative to the transcription start site and surrounded by GC-rich sequences. TATA-box binding protein (TBP), when in a TFIID complex, will bind to the TATA box. This is followed by the binding of TFIIA, TFIIB, TFIIF, RNA polymerase II, TFIIE, and lastly TFIIH for transcription initiation (Figure 14). The TATA box is the core promoter element in less than 50% of all human core promoters (Suzuki *et al.,* 2001).

TFIIB, besides facilitating TFIID-TATA box binding, can also bind to the TFIIB Recognition Element (BRE), which is found immediately upstream of the TATA box in about 12% of all TATA promoters (Lagrange *et al.,* 1998). Its consensus sequence is G/C- G/C-G/A-C-G-C-C followed by the 5' T of the TATA box. BRE may have either positive or negative effect on transcription (Evans *et al.,* 2001).

The initiator (Inr) element, when found, is located -2 to +4 relative to the transcription start site and has a consensus sequence of Py-Py(C)-A_{+1}-N-T/A-Py-Py in mammals (Corden *et al.,* 1980). Transcription is usually initiated at the A_{+1} nucleotide. This sequence is not bound by TBP but is bound by $TAF_{II}250$, which is a subunit of TFIID (Martinez *et al.,* 1994). Furthermore, although the TATA-Inr combination is known to be the best platform for RNA polymerase II binding, it

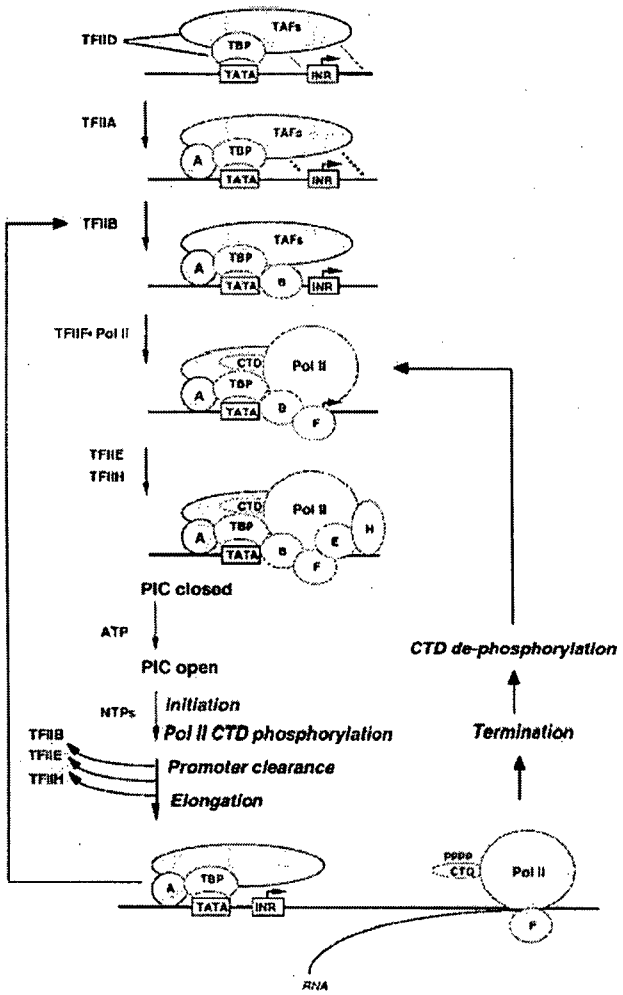


Figure 14. Model for a stepwise assembly and function of a pre-initiation complex (PIC). PIC assembly at a TATA-containing class II promoter is initiated by the binding of TFIID to the core promoter through both TBP interactions with the TATA box and TAF interactions with initiator (INR) or other downstream promoter elements. These interactions are stabilized by TFIIA. TFIIB further stabilizes the TBP-TATA complex and allows the recruitment of TFIIF, Pol II, TFIIE, and TFIIH in either a sequential manner (as shown) or as a preformed holoenzyme in which Pol II is in addition associated with the Mediator coactivator components (not shown in this figure). ATP-dependent promoter melting involves ATPase activities in TFIIH and is stimulated by THIIE to form the open complex that is competent to initiate transcription upon addition of ribonucleoside-triphosphate (NTPs). During initiation or early elongation the CTD domain of Pol II becomes phosphorylated by the CDK7 kinase activity of TFIIH allowing promoter escape/clearance and transcription elongation by hyperphosphorylated Pol II in association with TFIIF, while TFIIB, IIE, and IIH dissociate from the core promoter. After termination reinitiation might require de-phosphorylation of Pol II by a CTD-phosphatase. (Martinez E. Multi-protein complexes in eukaryotic gene transcription. Plant Mol Biol. 2002 Dec;50(6):925-47.)

has also been shown that, in the absence of a TATA box, Inr can cooperate with other TATA-less core promoter elements and direct accurate transcription initiation (Smale, 1997; Wieczorek *et al.*, 1998).

An example of a TATA-less promoter is the Downstream Promoter Element (DPE) (Burke *et al.*, 1996). It is located precisely at +28 to +33 relative to the transcription start site. The spatial distance for DPE-Inr promoters seems to be very important. For example, if the distance between DPR and Inr is altered, the core promoter activity is seriously reduced. Having a G nucleotide at position +24 is also preferred. On the other hand, the consensus sequence of DPE, A/G-G-A/T-C/T-G/A/C, is more variable than that of the TATA box. Like Inr, binding to DPE is accomplished by TFIID, not by TBP (Burke *et al.*, 1997; Kadonaga, 2002). DPE can be present as frequently as TATA box within promoter regions. A study on a database of 205 promoter regions with known transcription start sites in *Drosophila* has shown that 29% of the promoter regions contained TATA box only, 26% contained DPE only, and 14% contained both TATA box and DPE (Kutach *et al.*, 2000). Also, up to 31% of the core promoter regions contained no TATA box or DPE. Therefore, promoter elements other than TATA box and DPE are likely to exist.

Since both core promoter elements and RNA polymerase II are ubiquitously expressed and active, regulation is needed, which is achieved through interactions between regulatory DNA sequences and transcription factors.

1.9 Regulatory DNA Sequences

Regulatory DNA sequences, or transcription factor (tfactor) binding sites, are *cis*-acting, because they are located on the same DNA molecule that they regulate. Depending on their influence on transcription, tfactor binding sites can be organized into different groups:

1. Enhancers – positive regulatory elements that can increase basal transcription level over a long range (Blackwood *et al.,* 1998).
2. Silencers – negative regulatory elements that can decrease basal transcription level over a long range (Ogbourne *et al.,* 1998).
3. Insulators – neutral elements with a size of 0.5-3 kb that can constrain and define boundaries affected by enhancers or silencers (Geyer *et al.,* 2002).
4. Response Elements – flexible elements found within 1 kb upstream of the promoter that can either increase or decrease transcription based on external stimuli (Strachan *et al.,* 1999b).

Some enhancers like GC boxes and CCAAT boxes are integral to the core promoter, and they can be found in the vicinity of the transcription start site. For example, GC boxes (or Sp1 boxes), having a consensus sequence of GGGCGG, are usually found within 100 bp of the transcription start site, while CAAT boxes, having a consensus sequence of CCAAT, are usually found at position -75. On the other hand, cyclic AMP response element (CRE) can be found up to 200 bp upstream of the core promoter element. CRE has a consensus sequence of G-T-G-A-C-G-T-A/C-A-A/G. It can activate or deactivate transcription, depending on the conditions.

1.10 Transcription Factors

Unlike tfactor binding sites, tfactors are *trans*-acting factors because they are proteins encoded distantly and have to migrate to their sites of action. Each tfactor generally binds to a specific tfactor binding site through one of the common DNA-binding domains, which are structural motifs shared by different tfactors (Strachan *et al.,* 1999b). Some common motifs are:

1. Leucine Zipper - a leucine-rich helix that readily forms a dimer through coiled-coil interaction.
2. Helix-Loop-Helix (HLH) - related to the leucine zipper, and composed of one long and one short α-helix connected by a flexible loop. Two helices can pack against each other and permit both DNA binding and dimer formation.
3. Helix-Turn-Helix (HTH) - two short α-helices separated by a short amino acid that induces a turn.
4. Zinc Finger - a Zn ion bound by four conserved amino acids, usually four cysteine residues or two cysteine and two histidine residues.

Different tfactors have different mechanisms and effects on transcription. Currently, there are 2785 tfactor entries in the TRANSFAC database (Release 4.0). This number does not reflect the exact number of tfactors because of imperfect classification. Some of the best-studied ones are Simian-virus-40-protein-1 (Sp1), Activator Protein (AP-1), and CRE-binding protein (CREB).

1.10.1 Sp1

Sp1 was the first mammalian transcription factor for RNA polymerase II that was discovered and cloned (Briggs *et al.,* 1986; Kadonaga *et al.,* 1987). Having three two-cystein-two-histidine zinc finger motifs, it is also the founding member of a zinc finger tfactor protein family that includes

Sp2, Sp3, Sp4, basic transcription element B1 (BTEB1), and BTEB2. All Sp1-like tfactors tend to bind GC-rich sequences, like the GC box shown in Table 2 (Cook *et al.*, 1999). Because multiple Sp1 sites have been found near the core promoters of many housekeeping genes, Sp1 may be responsible for basal transcription (Lin *et al.,* 1996). Furthermore, multiple Sp1 sites can work synergistically to superactivate transcription. Sp1 activity is under tight control through phosphorylation and glycosylation and can upregulate genes for growth promotion and growth inhibition (Black *et al.,* 2001).

1.10.2 AP-1

AP-1 is not a single tfactor but instead represents a family of tfactors that possess leucine-zippers. Some members of the family are the Fos and Jun leucine zipper proto-oncogenes. It is related to cellular stress, UV irradiation, DNA damage, *etc.* Since leucine-zippers tend to dimerize through coiled-coil protein interaction, binding between these tfactors and DNA occurs at palindromic sequences (Wisdom, 1999). Different members of the family can form either homodimers or heterodimers. Originally thought to be stimulated by the phorbol ester 12-*O*-tetradecanoylphorbol-13-actate response element (TPA), AP-1 binds to the TPA-response element (TRE), which has a consensus sequence of 5' TGAGTCA 3' (Whitmarch *et al.,* 1996). For example, Jun-ATF-2 dimers can recognize 5' TGAGCTCA 3'. One of the characteristics of AP-1 is its ability to respond to an incredibly wide range of stimuli (*e.g.* cellular stress, DNA damage, oxidative stress, neuronal depolarization, T or B lymphocyte binding, cytoskeletal rearrangement, tumour necrosis factor α (TNFα), interferon-γ, ionizing and ultraviolet irradiation, MAP kinases). The two most potent inducers of AP-1 are UV irradiation, which leads to cell cycle arrest, and peptide growth factors (*e.g.* MAP kinase), which leads to cell cycle progression (Wisdom *et al.,* 1996).

Table 2. Examples of *cis*-acting elements recognized by ubiquitous transcription factors.

Cis element	DNA sequence is identical to, or a variant of	Associated *trans*-acting factors	Comments
GC box	GGGCGG	Sp1	Sp1 factor is ubiquitous
TATA box	TATAAA	TFIID	TFIIA binds to the TFIID-TATA box complex to stabilize it
CAAT Box	CCAAT	Many, *e.g.* C/EBP, CTF/NF1	Large family of *trans*-acting factors
CRE (cAMP response element)	CTGACGTA/CAA/G	CREB/ATF family, *e.g.* ATF-1	Genes activated in response to cAMP

From Strachan *et al.* 1999a.

1.10.3 CREB

CREB, a leucine zipper tfactor, is closely related to the cAMP pathway, which is stimulated mainly by hormones (Strachan *et al.,* 1999b). After a hormone receptor binds to a hormone, it activates the receptor-bound G protein, which then dissociates and activates the enzyme adenylate cyclase. This enzyme converts ATP to cAMP, which activates a wide variety of protein kinases, one of which is MAP kinase. All these protein kinases phosphorylate CREB protein, so that it can bind to an 8-bp cAMP responsive element (CRE), 5' TGACGTCA 3' (Whitmarsh *et al.,* 1996). Transcription is then activated. CREB can work with other tfactors like c-fos in the presence of growth factors (Ginty *et al.,* 1994).

1.10.4 Effect of Chromatin Structure on Tfactor Binding

When tfactors are activated, they can theoretically bind to all available tfactor binding sites in the genome and activate transcription in a wide array of genes. There must be control on tfactor-DNA binding to achieve transcription initiation specificity. This can be achieved in part through the regulation of chromatin structure.

When transcription is inactive, DNA is highly organized and compacted into chromatin. The most fundamental unit of chromatin packaging is the nucleosome, which consists of an octamer of core histones (H3, H4 tetramers, H2A dimer, and H2B dimer). Tighter or denser binding is achieved through the recruitment of H1 and H5 histone proteins. On the other hand, when DNA is transcriptionly active, it is less compact because the core histones are acteylated and H1 binding becomes weaker. Depending on the position in the cell cycle and the 3D orientation of the DNA, tfactor binding sites may be exposed (facing the surface) or hidden (facing the octamer) (Beato *et al.,* 1997). Furthermore, different chromatin sections may be opened or closed

depending on the presence or absence of other tfactors, thus achieving a fine tuning of transcriptional regulation.

1.10.5 Effect of DNA Methylation on Tfactor Binding

Besides chromatin structure, regulation of tfactor-DNA binding can be achieved through DNA methylation, which occurs through the action of DNA methyltransferase 1 at the 5 position of cytosine, converting a cytosine to a methylcytosine. Methylation occurs in about 4% of cytosines in the genome, and all methylcytosines are found in 5'CG 3' dinucleotides (Razin *et al.,* 1980). One would expect to find the majority of methylcytosine in CpG islands, which are regions with a high proportion of CG content that are more concentrated 5' instead of 3'of a gene (McClelland *et al.,* 1982). However, most CpG islands are unmethylated. Passive demethylation can occur during DNA replication, possibly under the effect of DNA binding factor (Wolffe *et al.,* 1999). Active demethylation involves demethylases, which are highly active in the developing embryo (Jost, 1993).

DNA methylation can interfere with tfactor binding in two ways (Attwood *et al.,* 2002). First, methylated DNA proximal to a promoter may recruit methylcytosine-binding proteins, which then recruit corepressor and histone deacteylases. This in turn changes the chromatin from an active to an inactive transcription state. Second, methylated CG dinucleotides may protrude into the major DNA groove and block the access of tfactors to their binding sites. AP-2, CREB, and Sp1 have been shown to be inhibited from binding to their corresponding recognition sequences by methylation (Comb *et al.,* 1990; Iguchi-Ariga *et al.,* 1989; Clark *et al.,* 1997).

1.11 *NF1* Gene Transcription

Studies using primer extension and RNAse protection on human melanoma and brain tissue culture cells have shown that the major transcription start site for *NF1* begins 484 nucleotides upstream of the translation start site (Wallace *et al.,* 1990; Harja *et al.,* 1994). There are alternate transcription start sites at +11 and -1 relative to the major transcription start site (Viskochil, 1998).

1.11.1 Potential Tfactor Binding Sites and Methylation

The 5' Untranslated Region (5' UTR) of *NF1* gene contains 5 potential AP-2 sites, 3 of which are 140-220 bp upstream of translation start site, and 2 potential Sp1 sites, located at 24 bp and 66 bp upstream of the translation start site (Harja *et al.,* 1994). Since AP-2 is restricted in expression with high abundance in neural crest cells (Mitchell *et al.,* 1991), it may play an important role in NF1, which involves tissues derived from neural crest, among many others. According to Hajra *et al.*, there are two potential AP2 sites at 623 bp and 650 bp upstream of the translation start site, two SP1 sites at -625 and - 649, one potential GT2 site at -636, one potential MT1 site at -584, one potential CREB site at -500, and one potential SRE site at -498, all within 200 bp upstream of the transcription start site (Harja *et al.,* 1994).

Some studies suggest that methylation may be a potential mechanism in regulating *NF1* gene expression (Mancini *et al.,* 1999). As mentioned, methylation can hinder tfactor-DNA binding. No methylation is observed in the vicinity of the Sp1 binding sites 625 bp and 649 bp upstream of the *NF1* translation start site. Furthermore, both the CREB and Sp1 sites upstream of translation start site can be blocked from binding to DNA by methylation. However, since no actual methylation of these binding sites is observed in neurofibrosacrcomas and neurofibromas,

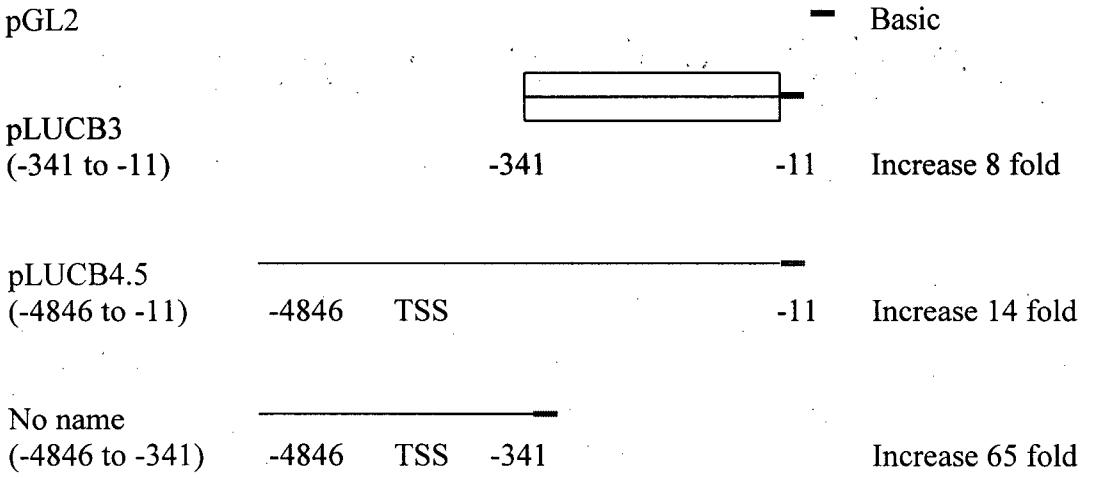
methylation may not be a major cause of NF1 gene inactivation during tumourogenesis (Luijten *et al.,* 2000). Furthermore, although methylation can be seen at -609, -429, -406, -383, -331, and -315 nucleotides relative to the transcription start sites in NF1-specific tumours, the other 64 potential CpG methylation sites near NF1 promoter region are unmethylated, and none of the methylated ones are located at predicted transcription factor binding sites.

1.11.2 Core Promoter Element

As mentioned, no TATA or CCAAT box exists in the *NF1* promoter region. Furthermore, no consensus sequence for Inr has been found (Viskochil, 1998). Two unpublished luciferase assay studies of *NF1* transcription activity have revealed some potential regulatory regions. Although both studies focused on regions with positive or negative effects on *NF1* transcription, these studies may provide an indication where the *NF1* core promoter element lies (Figure 15).

Using a luciferase assay, Purandare *et al.* have shown that the sequence between 4846 bp upstream and 11 bp upstream of the *NF1* translation start site can increase luciferase activity by 14 fold. The region between 341 bp and 11 bp upstream was able to act independently as a promoter. On the other hand, a construct that deleted the region between 11 bp and 341 bp upstream of the translation start site increased luciferase activity by 65 fold. These results suggest the presence of a promoter and a strong repressor in the region 11 bp to 341 bp upstream of the ORF. Furthermore, both predictions are located downstream of the transcription start site (Purandare *et al.,* 1996). Lastly, addition of more upstream sequence can increase luciferase activity, which may signify the influence of potential activating tfactors.

a) From Purandare *et al.*



b) From Rodenhiser *et al.*

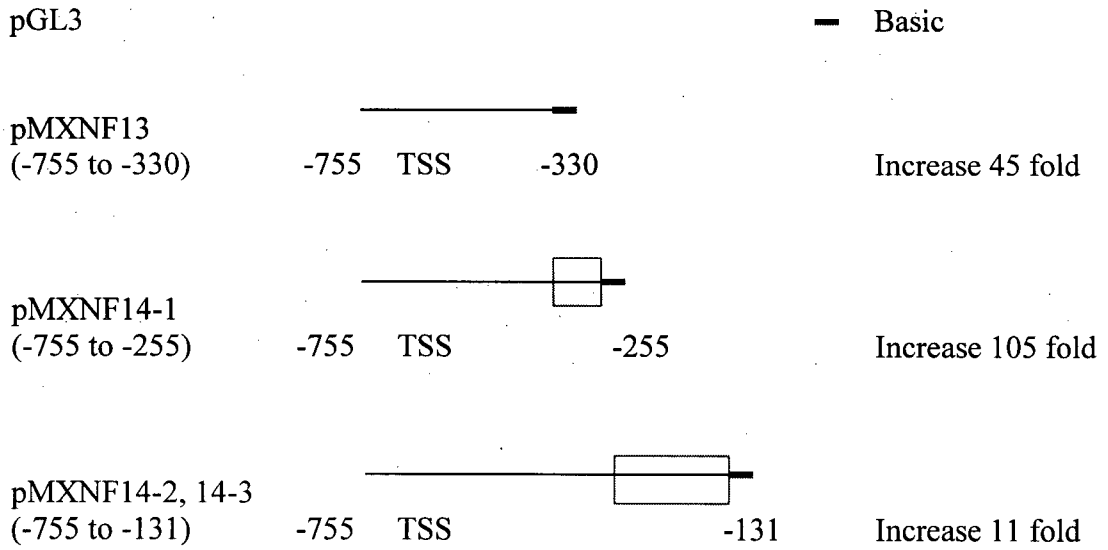


Figure 15. Summary of luciferase assay results from (a) Purandare *et al.*, and (b) Rodenhiser *et al.* Name, the location of the construct, and the luciferase activity are included. Locations in brackets are relative to the translation start site (+1), which is 484 bp downstream of the transcription start site (TSS). Note that the basic construct used in Purandare *et al.* is pGL2 (thick red line) while the one used in Rodenhiser *et al.* is pGL3 (thick green line). The increases in activity can be compared within the experiments but not across the experiments. Blue box (a) shows the region that may hold potential core promoter element, activator, and repressor. Pink box (b) shows the region that may hold potential core promoter element while green box (b) shows the location of possible repressor.

Rodenhiser et al. have performed similar luciferase assay experiments (Rodenhiser *et al.*, 2002). Different constructs were made that began at 755 upstream of the translation start site and ended at different positions downstream of the translation start site. The transcript with the most luciferase activity was pMXNF14-1, which includes the segment between –755 bp and -255 bp of the translation start site. Lengthening this construct to -131 bp or shortening it to -330 bp both decreased its effectiveness. These results suggest the presence of a repressor downstream of -255 and an activator, possibly a promoter element, located between -330 bp and -255 bp of the translation start site. Considering the two studies together, one may suggest a possible core promoter element between 255 bp and 341 bp upstream of the ORF.

1.12 Transcription Factor Prediction

There are two main methods for predicting potential tfactors that regulate gene expression. The first way is through coexpression. The second way is through direct tfactor binding site detection.

1.12.1 Coexpression

Any given cellular action usually involves more than one protein or gene product. Therefore, organisms often have to orchestrate transcription for different genes together. A simple way to do so is to regulate genes with related functions with common tfactors. By collecting mRNA and protein expression data, researchers can often associate actions of different genes together to form a gene network. If one of the genes is known to be regulated by certain tfactors, there is a possibility that those tfactors will affect other genes within the same network. Furthermore, if there are known tfactor binding sites for a particular gene, researchers may be able to discover binding sites for those tfactors for the other coexpressed genes by comparing their sequences.

There are computer programs and analytical tools designed to facilitate the investigation of coexpression among different genes (Gasch *et al.*, 2002). Unfortunately, the *NF1* gene is ubiquitously expressed. There is no coexpression data that will be useful for detecting tfactors for *NF1* in this study.

1.12.2 Direct Tfactor Binding Site Prediction

Potential tfactor binding sites can be predicted using the powerful tools of bioinformatics (Kanehisa *et al.*, 2003). Tfactor binding site prediction can be achieved by analysing the DNA sequence of a gene for known consensus tfactor binding site sequences, which are usually very short. However, unlike restriction endonucleases that have very specific recognition sequences, tfactor binding sites are often degenerate. Minor changes in a tfactor binding site may cause decreased affinity rather than abolishing tfactor-DNA binding (Roulet *et al.*, 1998). Therefore, a tfactor may bind to different DNA sequences. This limits the effectiveness of predicting tfactor binding sites solely based on a perfect match in DNA sequence.

This problem is alleviated in part by using a weight matrix, which was first used to characterize *E. coli* transcription and translation initiation sites (Harr *et al.*, 1983; Stormo *et al.*, 1982). A weight matrix is a two-dimensional table in which each row corresponds to one of the four letters of the DNA alphabet and each column corresponds to consecutive positions of a tfactor binding site sequence. A different nucleotide at each position will be given a different number, which reflects a tfactor's binding strength at that position when recognizing the binding site. Each tfactor will have a different weight matrix, which is obtained by compiling sequence variations of the naturally occurring binding site of that tfactor, recording the differences in affinities of different sequences, and calculating the effect of different nucleotides at each position. Researchers can then assign a final score to any given sequence by combining all the

corresponding numbers in the matrix for that sequence. Using a set threshold, researchers can classify the sequence as either nonfunctional or a potential binding site for the tfactor (Frech *et al.*, 1997). TRANSFAC® is the biggest collection (library) of tfactor data that is available publicly (Matys *et al.*, 2003). The most current version, Release 6.0, contains 6627 entries for putative tfactor binding sites for different eukaryotic genes, with species ranging from yeast to human.

Detection of tfactors by weight matrices has several weaknesses:

1. Because of the need to rely on TRANSFAC® or another library, no novel tfactors can be predicted.
2. Some tfactor binding sites have no natural variants, so they do no have a weight matrix.
3. Biologically important tfactor-DNA binding can be very tolerant so that even weak interactions may be very important. Setting too high a cutoff for weight matrices will act against the prediction of these tfactors. Similarly, values above the cutoff may not represent true tfactor binding sites.

There are many prediction programs for 'potential' tfactor binding sites that are based on weight matrices and TRANSFAC libraries. The details of the programs used in this study can be found in the Methods section. The adjective 'potential' cannot be stressed enough because the actual function of each prediction has to be validated through experiment. Furthermore, these programs often generate predictions that do not agree with one another, and many of the predictions are false-positive. Since all of these programs use the same TRANSFAC library and similar algorithms, using multiple programs does not necessarily alleviate the problem. This is where phylogenetic footprinting becomes important.

1.13 Phylogenetic Footprinting

Evolution explains the two principles of life – diversity and unity. Diversity refers to the wide array of different life forms on the earth. Unity refers to the common biochemical, cellular, genetic, and physiological characteristics shared by different organisms. The four basic nucleotides in DNA, the amino acids used in proteins, the electron transport chain in respiration, and organelles in the cells, all reflect the common origin shared by different organisms on the earth. It is, therefore, not surprising that many genes, including *NF1*, are shared by different organisms. Because the functions of the protein product are determined by the genetic sequence, it is expected that exons, which are the coding sequences of proteins, would be highly conserved across different species through evolution. For a long time, however, the importance of introns was overlooked because of their non-coding nature.

The basic principles of evolution and natural selection suggest that changes in functional DNA sequences are generally selected against because random mutations are usually deleterious. Coding regions between humans and mice, for example, share an average homology of 85% in their DNA sequences. On the other hand, non-coding introns share an average homology of 69%. The homologies of 5' UTR and 3' UTR are 75-76%, with similar degree of homology in the first 200 bp upstream of the transcription start site (Waterson *et al.*, 2002). Since these regions upstream of the transcription site have a higher homology than non-coding regions in general, they are likely to be functional and may contain potential transcriptional control elements like core promoter regions and tfactor binding sites. However, homology comparison alone cannot predict what functions areas of interest may have. Tfactor detection programs and phylogenetic footprinting can, therefore, be complementary to each other. For example, a

combined approach has been used in the search for potential *cis*-regulatory elements in the promoter of 5-Lipoxygenase in humans and mice (Silverman *et al.*, 2002).

Besides *Homo sapiens*, genomes from *Caenorhabditis elegans*, *Drosophila melanogaster*, *Fugu rubripes*, *Anopheles gambiae*, *Mus musculus, Rattus norvegicus*, *Caenorhabditis briggsae*, *Saccharomyces cerevisiae*, and many other eukaryotic organisms are completely or close to completely sequenced. With the ever-increasing number of eukaryotic genomes becoming available, bioinformatics for comparative study becomes feasible (Ureta *et al.*, 2003). Many computer programs for aligning genomes and detecting regulatory regions have been developed, and most of them are available free for academic use.

The first comparative study of *NF1* for potential regulatory regions based on human and mouse sequence was done in 1994 (Hajra *et al*, 1994). This study demonstrated highly conserved regions and several conserved tfactor binding sites within 1000 bp upstream of the translation start site of humans and mice. Another study attempted to align flanking genes of *Fugu rubripes NF1* to human *NF1*, but those genes are located very far away or on different chromosomes in humans (Kehrer-Sawatzki *et al.*, 2000). Since then, there is no entry in PubMed concerning comparative study of human *NF1* with other organisms.

1.13.1 Special Note on *Fugu rubripes*

Fugu rubripes is the poisonous and delicious Japanese pufferfish. The International *Fugu* Sequencing Consortium has recently completed the draft assembly of 12,403 contigs covering approximately 90% of the genome (Aparicio *et al.*, 2002). The *Fugu* genome, while possessing a similar gene repertoire to mammalian genomes, has a total size of less than 400 Mb, which is eight times smaller than mammalian genomes. There is much less non-coding sequence in *Fugu*.

Therefore, comparative study between the human and *Fugu* genome can potentially pinpoint functionally important segments of intronic regions (Aparicio *et al.*, 2002). With an evolutionary distance of 450 million years between humans and *Fugu*, compared to 40.5 million years between humans and mice or rats (Kumar *et al.*, 1998), regions that are highly conserved between humans and *Fugu* may contain important regulatory elements. Comparative studies between human and *Fugu* have been done on several genes (Goode *et al.*, 2003; Annilo *et al.*, 2003).

1.14 Thesis Rationale and Objectives

The 1994 comparative study of the *NF1* gene was limited to 1000 bp upstream of the transcription start site in human and mouse (Hajra *et al*, 1994). With many eukaryotic genome sequencing projects completed or close to completion and the increased knowledge on transcription factor binding sites, there is a need to conduct another comparative study. Therefore, I studied the 5' Upstream Region (5UR) and Intron 1 of the *NF1* gene from four vertebrates - human, mouse, rat, and pufferfish. This study covered a larger region with updated sequences.

There were two main objectives for this study. First, phylogenetic footprinting was done on the 5' Upstream Region (5UR) and Exon 1/Intron 1 (EI1) regions of the *NF1* gene in the four species to identify and characterize the most homologous regions. The second objective was to uncover potential regulatory regions based on sequence alignment together with transcription factor binding site and promoter detection programs.

Chapter 2. Methods

2.1 Overview

The success of this research depends on accurate and current genetic sequences. The human, mouse, rat, and *Fugu NF1* gene cDNA, 5UR, and EI1 sequences, as well as the neurofibromin amino acid sequences were first located and downloaded from UCSC, NCBI, or ENSEMBL. The degree of cDNA and protein homology was then calculated. Next, 5UR and EI1 regions were compared using the sequence alignment program mVISTA, followed by analysis using a Perl program called Frameslider that was written to calculate identity. Highly homologous regions were located based on cDNA homology. Regions surrounding these highly homologous regions were then analyzed using tfactor binding site prediction programs. Predictions from different organisms were compared, and ones that were shared by human and at least one other species at the aligned positions were selected. Lastly, all data were summarized and presented using a graphic program.

2.2 Introduction to Programs Used for Analysis

Many different programs have been used for this bioinformatics study. In order to understand the research logic, a basic understanding of the functions of the different programs is necessary. They can be broken down into five categories:

1. Sequence Search and Homology
2. Sequence Alignment and Homology
3. Promoter Search

4. Tfactor Search

5. Graphics Display

Note that there are programs that perform alignment together with tfactor binding site prediction (*e.g.* ConSite). However, for most of these programs, either the output format is hard to maneuver or some information like exact nucleotide location is lost. Therefore, alignment and tfactor binding site prediction were done separately.

2.3 Sequence Search and Homology

2.3.1 BLAST (http://www.ncbi.nlm.nih.gov/BLAST/.)

BLAST (Basic Local Alignment Search Tool) is a set of similarity search programs designed in 1990 (Altschul *et al.,* 1990). When a user supplies a query, whether it is a sequence of DNA or protein, BLAST can search for similar sequences from the available databases in NCBI. The BLAST algorithm finds similar sequences by building an index or dictionary of short subsequences called words for both queries and the database. The program then searches for exact matches by comparing words in the query to words in database. There are many different variations of BLAST (Table 3). A user can search a subset within a database (*e.g.* RNAs, ESTs, Protein) and adjust different parameters for different alignment stringencies. The two most important parameters are word size and Expect value. Word size reflects the length of individual words in the indices for the query and database. Word size can range from 7 to 11. Increasing the word length increases the stringency. The Expect value reflects the number of matches expected between the query and database by chance alone. The lower the Expect value,

Table 3. BLAST programs.

Program	Description
blastp	compares an amino acid query sequence against a protein sequence database
blastn	compares a nucleotide query sequence against a nucleotide sequence database
blastx	compares a nucleotide query sequence translated in all reading frames against a protein sequence database.
tblastn	compares a protein sequence against a nucleotide sequence database dynamically translated in all reading frames.
tblastx	compares the six-frame translation of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database.

From NCBI BLAST Tutorial
(http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/query_tutorial.html)

the more stringent the test. If the statistical significance assigned to a match is above the Expect value, that match will not be reported. Similarly, matches with lower Expect value are more significant statistically. The main BLAST program used in this research was blastn with default settings for organism-specific genomic BLAST.

The main limitations of BLAST are its speed and output. Analysis can be time consuming, especially under low stringency settings. Furthermore, although the output provides useful statistics and the database (accession number) where matches are located, it lacks a convenient coordinate system that enables the user to the pinpoint location of matches in the genome without doing additional searches.

2.3.2 Pairwise BLAST (http://www.ncbi.nlm.nih.gov/blast/bl2seq/bl2.html)

This is another variation of BLAST that was very useful in this study. Using the same BLAST engine, local alignments are found by comparing two user-supplied sequences (Tatusova *et al.*, 1999). Both blastn and blastp are used. The former is useful in aligning and locating a specific query sequence relative to the other sequence. The latter is useful in comparing the degree of protein homology. Unless specified, default settings were used in this study.

2.3.3 BLAT (http://genome.ucsc.edu/cgi-bin/hgBlat?command=start&org=human)

BLAT is short for BLAST-Like Alignment Tool (Kent, 2002). Although it is spelled similarly, BLAT is not BLAST. Unlike BLAST, BLAT builds an index of the database (but not the sequence) and searches linearly along the sequence (instead of the database). For example, the database used in BLAT is broken down into non-overlapping 11-mers for DNA and 4-mers for protein. Furthermore, BLAT output stitches together different alignments to form one big alignment, instead of producing individual broken local alignments as in BLAST. Overall,

BLAT is much quicker than BLAST and can give a clearer picture of alignment. It is used extensively in the UCSC database as an extremely efficient tool in pinpointing the exact location of a specific query sequence within a specific genome database (*e.g.* chromosome number and the coordinates). However, BLAT does not have any index for *Fugu*, so a *Fugu* BLAT search is not currently available. Also, there is an upper limit on the query sequences of 25000 nucleotides in DNA, which is less than the length of *NF1* human 5UR or intron 1 (more details on this later). It was, therefore, necessary to use other alignment tools as well.

2.4 Sequence Alignment and Homology

2.4.1 mVISTA (http://www-gsd.lbl.gov/vista/)

mVISTA stands for main VISualization Tool for Alignment, which is designed for comparative genomics (Mayor *et al.*, 2000). The user can input two or more sequences for several one-on-one alignments simultaneously. For example, if human (H) sequence is input as the base organism, and comparisons to sequences of mouse (M), rat (R), and pufferfish (F) are desired, the program returns alignments of HvsM, HvsR, and HvsF. Note that if only three organisms are picked, for example, H, M, and R, there is no need to specify a base organism and all possible comparisons are provided (*e.g.* HvsM, HvsR, and MvsR). Alignment is achieved by sliding a window of predefined length along the comparison sequences and searching for regions of high homology. In this study, alignment was performed under the setting "one-cell organism". This setting was used to avoid sequence masking for repeat regions, which may contain tfactor binding sites. The window size and conservation level for the identity calculation do not affect the alignment itself, so default settings were used.

The advantages of mVISTA are that multiple large sequences can be aligned quickly and the memory requirement is low. However, mVista alignment is based on the assumption that the order of conserved regions (synteny) will be preserved through evolution. Therefore, mVISTA is not suitable for genome-wide alignment, but it can be used for smaller local alignment of different regions of the *NF1* gene.

Another problem with mVISTA is its output. Although mVISTA uses sliding windows in its alignment, the output does not define the location of individual windows with respect to the genome as a whole. Final identity scores are given along the alignment, but it is impossible to pinpoint the exact coordinates that correspond with a particular score. The scores for individual windows of a fixed size are also not available. This is an important limitation if a nucleotide by nucleotide comparison is being performed. For this project, there was a need to tap into the hidden information available in the alignment so that the identity score of individual windows could be determined and appropriate regions selected for further analysis.

2.4.2 Frameslider

Frameslider (frameslider.pl) is a Perl program that I designed to report percent identity of individual window comparisons along an alignment between two species (Appendix 1). First, the user creates two text files by converting an mVISTA alignment from horizontal format to vertical format supplied with an index number (Figure 16). This can be done easily using EXCEL and Word Cut-and-Paste, Replace, and Data-to-column commands. Also, the two files must be named with asequence.txt as the first strand and bsequence.txt as the second strand in the same directory as frameslider.pl. Executing Frameslider will prompt the user to input the window size needed for comparison. Suppose a window size of 5 is picked (Figure 16). The program starts at nucleotide 1 on the first strand and compares this with nucleotide 1 on the

a) Hypothetical Alignment:

```
Human ATGCATGCA-GGTTGC----
      ||||| ||| |  |||
Mouse ATGCA-GCATGCATGCAAAA
```

b) Input Format

asequence.txt

1	a
2	t
3	g
4	c
5	a
6	t
7	g
8	c
9	a
10	-
11	g
12	c
13	a
14	t
15	g
16	c
17	-
18	-
19	-
20	-

bsequence.txt

1	a
2	t
3	g
4	c
5	a
6	-
7	g
8	c
9	a
10	t
11	g
12	c
13	a
14	t
15	g
16	c
17	a
18	a
19	a
20	a

c) Output Format

Page 1

Start	End	ANucleotide	BNucleotide	Homology
1	5	a	a	1
2	6	t	t	0.8
3	7	g	g	0.8
4	8	c	c	0.8
5	9	a	a	0.8
6	11	t	-	0.8
7	12	g	g	0.8
8	13	c	c	0.6
9	14	a	a	0.6
10	16	-	t	0.6

Figure 16. Demonstration of Frameslider with window size of 5. a) alignment output from mVISTA. b) input file format. c) sample output. Note that frameslider skips gaps (character '-') to obtain a correct window size, and, therefore, skips one comparison if a gap occurs on the first strand. It will not skip a gap on the second strand and will take it into the identity calculation.

second strand. If they are identical, the number of identical nucleotides will increase by 1; otherwise, it will remain zero. Since the number of nucleotides being looked at is not equal to the window size (5 in this example), the program moves to the next nucleotide on both strands and repeats the analysis. Any gap on the first strand will be ignored during window size calculation, and no comparison will be made at that position. This ensures identity calculation will include the same number of nucleotides. On the other hand, a gap on the second strand will count as a mismatch in the identity calculation. When comparison of five nucleotides has been completed, the programme will report the average identity (*e.g.* the total number of matches divided by the window size). The output.txt is updated after each window is compared, and the identity score is reset to zero. The program then moves to the next nucleotide as the starting point and begins the next analysis. The identities of different windows can be ranked very easily in EXCEL by using the 'Sort' command and precise information on nucleotide coordinates is provided.

In this study, a window size of 50 was chosen between HvsM, HvsR, and MvsR alignments because this size has been shown to be optimal for the comparison between human and mouse genomes (Waterson *et al.*, 2002). A window size of 30 or 20 was used for alignments against *Fugu* because of the larger evolutionary distance. Windows with the highest identity were investigated further.

Frameslider has two main weaknesses. First, a misleading result can occur if there are extensive gaps in the top strand. For example, an alignment like this will report 100% identity when in fact the alignment is likely to be incorrect:

```
Human  AAAAA--------------AAAAAAAA------------AAAAAAA
       |||||              ||||||||            |||||||
Mouse  AAAAACCCCCCCCCCCCCCCAAAAAAAACCCCCCCCCCCCCAAAAAAA
```

This problem can be spotted quite easily during manual data analysis. Furthermore, reversing the order of two strands demonstrates this artifact.

The second problem is harder to solve. Because Frameslider is based on the mVISTA alignment, Frameslider cannot compensate for the limitations of mVISTA. Erroneous alignments can occur in mVista when the sequences under investigation are too different in length or too far apart in their evolutionary history. Sequence length can be chosen wisely, but alignment of sequences that are evolutionarily distant may be better done with BLAST or pairwise BLAST, where the parameters can be easily adjusted and no assumptions on conserved order for homologous regions is necessary.

The code for Frameslider can be found in Appendix 1.

2.5 RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker/)

RepeatMasker is a program written by Arian Smit (Smit *et al.,* 1996) that screens a DNA sequence against a library of repetitive elements like interspersed repeats or low complexity DNA sequences. The output returns a masked query sequence with all repeats replaced by 'N's and a table annotating the exact location of repeats. Almost 50% of the human genomic sequence is masked by the program. The comparison of DNA sequences to the repeat library is performed by a program called cross_match, an efficient implementation of the Smith-Waterman-Gotoh algorithm developed by Phil Green. Theoretically, there is no sequence size

limit for this program. However, depending on the sensitivity setting, if the input sequence exceeds 10kb, the connection may time out if the online version is used. For the purpose of this study, all queries were done in slow speed, high sensitivity setting, with the appropriate organism DNA source. Since the sequence is not masked before alignment to avoid losing potential tfactor binding site regions, only the annotation table was selected as output.

2.6 Promoter Identification

2.6.1 GenomatixSuite (http://www.genomatix.de)

GenomatixSuite is a collection of bioinformatics tools designed for RNA polymerase II promoter and tfactor binding site prediction (*e.g.* PromoterInspector, El Dorado, Chip2Promoter, MatInspector, *etc*). An academic license is available without cost, but there are limitations on the number of analyses that can be done and other restrictions.

The user can import a sequence to search for potential promoters in PromoterInspector (Scherf *et al.*,

2000). PromoterInspector predicts promoter regions rather than the core promoter. It looks for a combination of different components of a promoter region separated by an acceptable number of wildcard nucleotides 'N'. The acceptable range is derived from experiments. The prediction is checked against three classifiers, which are exons, introns, and 3' UTRs. Each classifier consists of a collection of training sequences. Each prediction is compared to promoter training sequences and one of the classifiers. If more hits are found within the promoter training sequence database, the sequence is assigned as a potential promoter prediction. If the sequence passes all three classifier tests, then the prediction is output as a valid promoter prediction. The

program claims to have a specificity of 85% but a sensitivity of 48%, so it misses promoters half of the time. The limitation on length input is 100000 bp, which means the whole *NF1* gene cannot be input at one time for this analysis.

GenomatixSuite also contains two other programs to search for promoters and other additional features (*e.g.* exon-intron boundary, gene annotation) but the user must have the mRNA sequence of the gene of interest. The user can enter the mRNA sequence directly through El Dorado or the NCBI accession number of the mRNA through Chip2Promoter. Both programs then extract additional information about the gene associated with this mRNA from a copy of the NCBI database stored within GenomatixSuite. Annotated features like exon-intron boundaries are displayed. Furthermore, PromoterInspector predictions are performed on regions adjacent to the mRNA sequences. If there is an annotated promoter, then the program will compare the PromoterInspector predictions and the annotations to see whether they agree with each other. Another program in GenomatixSuite is MatInspector, which will be discussed below in the Tfactor Prediction section.

One main drawback of El Dorado and Chip2Promoter is that they use build 31 of the human genome from NCBI as the database rather than the most current NCBI database. Build 31 was released on Nov 15, 2002. Build 33, which was used in the other analyses in this study, was released April 10, 2003, so the database used by GenomatixSuite is outdated.

2.6.2 Dragon Promoter Finder

Dragon Promoter Finder (DPF) is a relatively new program available on the internet without charge for academic users. DPF is designed to locate promoter regions based on Transcription Start Site (TSS) prediction obtained from five independent promoter recognition models. A

sliding window comparison is performed for each model to look for promoters, exons, and introns by weight matrices. This program claims to perform better than PromoterInspector (Bajic *et al.*, 2002), and its main attraction is higher sensitivity (up to 80%), compared to the 48% offered by PromoterInspector. However, DPF requires two precautions. It can only take 10,000 bp for analysis, which is much shorter than the human *NF1* gene. In addition, because of the algorithm used, DPF cannot make any prediction for the promoter region if the TSS is located within the first 150-200 nucleotides (depending on setting) or the last 49 nucleotides of a sequence.

2.7 Tfactor Prediction

2.7.1 MATCH™ (http://www.gene-regulation.com/cgi-bin/pub/programs/match/bin/match.cgi)

There is no full publication on MATCH™, but a poster was presented at the German Conference on Bioinformatics in 2001 (available at http://www.bioinfo.de/isb/gcb01/poster/index.html). BIOBASE is the creator of this tool. MATCH™ relies on two scores for tfactor prediction: a core-similarity score and a matrix-similarity score; both range from 0 to 1, with 1 for an exact match. The core-similarity score weighs the quality of a match between the test sequence and the core sequence of a matrix, which consists of the five most conservative positions in a matrix. The matrix similarity score determines the quality of a match between the test sequence and the matrix. Because the normally time-intensive martix-similarity calculation is only performed for matches above a certain cut-off value in core-similarity score, the core-similarity score can act as a screen for tfactor prediction. This feature can be turned off easily by setting the cut-off score as zero.

The best features of MATCH^{TM} for this study are the three pre-calculated cut-off value settings. They are:

1. Minimize false-negative (minFN)
2. Minimize false-positive (minFP)
3. Minimize the sum of both errors (minSUM)

To decide cut-off values for the minFN setting, the programmers applied no cut-off values for initial predictions for their experiment sequences. Then cut-off values were applied to screen out a maximum 10% of the predictions. When these cut-off values are applied, detection can approach 90% sensitivity (Pickert *et al.*, 1998). Specificity is compromised, so there will be large number of false-positives, and the computing time is great with this setting.

The second setting is opposite to the first setting because its specificity approaches 99% with a loss of sensitivity (Pickert *et al.*, 1998). Cut-off value settings for minFP are generated from experiments with exon 2 and 3 sequences, assuming they should have no biologically active tfactor binding sites.

The minSUM method is based on a dissertation written in Germany (Reuter, 2000). Using exon sequences, the programmers calculated the number of matches by first setting cut-off values for minFN10 (identical to minFN) and defining the result as 10% FN and 100% FP. Subsequently, different cut-off values were used on the same sequence to obtain different number of matches with different FN and FP. The cut-off values that yielded the minimum sum of FN and FP were then used for the minSUM setting.

2.7.2 MatInspector

MatInspector's large library (>200) of tfactor binding site matrices was created with MatInd, a program for matrix creation (Quandt *et al.,* 1995). This program uses published matrices as entries with an emphasis on sequences with experimentally verified binding capacity as well as the TRANSFAC database.

Similar to MATCH™, MatInspector allows the user to define core and matrix similarity (i.e. cut-off values). However, the core used in MatInspector is usually 4 bp long instead of 5 bp in MATCH™. Although MatInspector does not have minFN, minFP, or minSUM settings, it has a powerful feature that allows the user to select optimized matrix similarity thresholds for each individual matrix. A drawback is the slight inconvenience of excluding the common name for tfactors in the output and limitation on the number of uses with the free academic license. For the purpose of this study, the optimized setting for matrix similarity and core similarities of 0.70 (lowest) and 1.00 (highest) were used.

2.8 Graphic Display

2.8.1 GFF

GFF, Gene-Finding Format or General Feature Format, is a special format in a data sheet containing genomic information (Figure 17). One genomic feature occupies one line. The details of the feature, which are the sequence name, source, feature, starting nucleotide, ending nucleotide, score, strand, frame, attributes, and comments, are entered on the same line, separated by a tab with no extra space. This rule must be observed in order for the data to be read correctly. Using GFF allows genomic features to be summarized in a standard way without

SEQ1	EMBL	atg	103	105	.	+	0
SEQ1	EMBL	exon	103	172	.	+	0
SEQ1	EMBL	splice5	172	173	.	+	.
SEQ1	netgene	splice5	172	173	0.94	+	.
SEQ1	genie	sp5-20	163	182	2.3	+	.
SEQ1	genie	sp5-10	168	177	2.1	+	.
SEQ2	grail	ATG	17	19	2.1	-	0

Figure 17. Example of GFF. Columns from left to right are sequence name, source, feature, starting nucleotides, ending nucleotides, score, strand, and frame. Additional columns for attributes and comments can also be input. Note that each column is separated by a <Tab> character but not by a <Space>.

loss of information. Furthermore, a standardized format allows researchers to interpret data with different computer programs without generating different data files for different programs. More information on GFF can be found on the Sanger Centre Website (http://www.sanger.ac.uk/Software/formats/GFF/).

2.8.2 Sockeye (http://www.bcgsc.ca/gc/bomge/sockeye)

Created by BC Genome Sequencing Centre (BCGSC), Sockeye is a Java application capable of displaying genomic information in a 3-dimensional environment. The program can be run on Windows or Linux platforms. Users can visualize large quantities of genetic annotation on a DNA strand by supplying data in GFF. Also, users can select an individual feature and obtain its detailed information. Since multiple 'tracks' (*e.g.* for different organisms) can be displayed at the same time, this is an excellent tool for comparative studies. Sockeye-generated graphics can be saved in JPEG format.

This program does have some limitations. For example, unless used on a very powerful computer, a user cannot enter more than 1000 features. Therefore, it may not be possible to display the degree of homology along two DNA sequences, depending on the length of the sequence. Furthermore, because the program is still in development, there are some minor bugs. Nevertheless, this is one of the best graphics program for displaying genomic information.

2.9 Data Sources and Version

2.9.1 Sequence Database *NF1* gene

The most current data available on each species were used. For human, mouse, and rat, the whole *NF1* gene and its 5' Upstream Region (5UR) sequences were obtained from UCSC (http://genome.ucsc.edu/). The versions are April '03, February '03, and January '03, respectively. For pufferfish, the most recently updated NCBI contigs were used to cover different regions. The 5UR region to intron 2 was covered by CAAB01003481 (updated June 02), exon 3 to part of intron 4c was covered by AF197897 (updated July 00), part of intron 4c to intron 9 was covered by AC064564 (updated August 99), and the rest of the gene was covered by CAAB01003123 (updated July 02). The fruitfly *NF1* gene sequence was obtained in ENSEMBL (www.ensembl.org; version 14.3.1 updated 3 March 2003) at chr3R:21798100-21810826.

2.9.2 TRANSFAC

The TRANSFAC database was created by Wingender *et al.* (2000). Access is free. The database can be found at http://transfac.gbf.de/TRANSFAC/. It is the biggest collection of eukaryotic transcription binding site matrices available. The database is divided into different classes – SITE, GENE, FACTOR, CELL, CLASS, and MATRIX. SITE gives information on individual (putatively) regulatory sites within eukaryotic genes. GENE gives a short description of the gene in which a site (or group of sites) belongs to, including the gene(s) that are under the influence of that particular tfactor and/or the gene that codes for that tfactor. FACTOR describes the proteins binding to these sites. CELL gives brief information about the cellular source of proteins (*e.g.* cell lines, tissues, or organs) that have been shown to interact with the sites.

CLASS contains some background information about the transcription factor classes, while the MATRIX table gives nucleotide distribution matrices for the binding sites, if available.

SITE is the most important class for this research. The newest version of TRANSFAC is Release 6.0, and there are 6627 entries under SITE. Both MATCH^{TM} and MatInspector use the TRANSFAC database.

2.9.3 Eurkaryotic Promoter Database (EPD)

Eukaryotic Promoter Database was created by Praz *et al.* (2002). Access is free. The database can be found at http://www.epd.isb-sib.ch/. It is an annotated, non-redundant collection of eukaryotic RNA polymerase II promoter regions for experimentally-determined transcription start sites.

The most current version is Release 74. The database contains 2994 promoter region sequences that are available for download. Each promoter region is defined as 499 bp upstream and 100 bp downstream of a transcription start site. This definition is fixed and cannot be adjusted.

2.9.4 Transcription Regulatory Regions Database (TRRD) – TRRDUNITS

The Transcription Regulatory Regions Database was created by Kolchanov *et al.* (2002). Access is free. The database can be found at http://wwwmgs.bionet.nsc.ru/mgs/gnw/trrd/. Users can also input a sequence and BLAST against the data set. However, the exact location of the alignment within a genome is not provided in the academic version, which is a severe limitation for this study.

TRRD contains only experimental data, which are annotated by literature citations. TRRD release 6.0 includes the information on 1167 genes, 5537 transcription factor binding sites, 1714 regulatory regions, 14 locus control regions and 5335 expression patterns obtained through 3898 scientific papers. This information is arranged in seven databases: TRRDGENES (general gene description), TRRDLCR (locus control regions); TRRDUNITS (regulatory regions: promoters, enhancers, silencers, *etc.*), TRRDSITES (transcription factor binding sites), TRRDFACTORS (transcription factors), TRRDEXP (expression patterns) and TRRDBIB (experimental publications). TRRDUNITS was used in this study because it contains sequences for promoters, enhancers, and silencers. There were 1967 entries in this database as of March 18, 2003.

2.9.5 Rfam (http://rfam.wustl.edu/)

Rfam, the RNA family database, is a collection of sequence alignments and covariance models representing non-coding RNA families. The current (June 2003) release contains 165 models, and each model has a consensus base-paired secondary structure. The user can input a DNA sequence, and Rfam will look for potential RNA motifs by combining a BLAST and covariance model search (Griffiths-Jones *et al.*, 2003). The current limit on the size query is 2 kb.

2.9.6 SCOR (http://scor.lbl.gov/index.html)

SCOR is a collection of three-dimensional RNA structures. The current release (July 2002) contains 261 RNA structures from Protein Data Bank, 402 internal loops (two helices connected by two strands), and 295 external loops (one helix capped by one RNA strand). Tertiary interactions including helix-helix interactions, loop-loop interactions, A-minor interactions, loop or helix-loop interactions, pseudoknots, and tetraloop-tetraloop receptor interactions are also being classified (Klosterman *et al.*, 2002). The user can input a sequence, and SCOR will look for an exact match to any of the known tertiary structures in its database. Alternatively, a user

can use special characters in a sequence string to allow more flexibility in the search (http://scor.lbl.gov/help/sqsearchHelp.html). Although this feature is useful, it becomes increasingly cumbersome as a search string gets longer. This rigidity is an important limitation of SCOR searches.

Chapter 3: Experimental Results

3.1 Transcription Start Site (TSS)

The *NF1* gene has been reported to have alternative TSSs, with the major one at 484 bp upstream of the translation start site as determined by primer extension assay and RNase protection assay (Marchuk *et al.*, 1991; Harja *et al.*, 1994). Two other TSSs are located at +11 bp and -1 bp relative to the major TSS (Viskochil, 1998). In order to search for additional TSSs, another search on the current human EST database was done (NCBI human genome build 31 released April 10, 2003).

Using BLAT in UCSC, NCBI's entry of NF1 mRNA (NM_000267, updated 03-APR-2003) was found to begin at nucleotide 29272015 on human chromosome 17 (Hchr17:29272015). Compared to the ORF, which begins at Hchr17:29272226, the TSS was 211 bp upstream of the translation start site if the beginning of this mRNA is assumed to be a TSS. Next, this mRNA was compared with the genomic EST database using blastn. Four EST clones AA832486, AA837064, AA489674, and AA807249 were found. However, these clones were actually in the middle of the *NF1* gene instead of the beginning of the gene, which seemed to have no EST collection in the NCBI database. Because this study requires a consistent coordinate system for the genomic sequence, and there seems to be a range of possible TSSs, all sequences were labeled in relationship to the translation start site at Hchr17:29272226.

3.2 Promoter Region and Core Promoter Element

3.2.1 GenomatixSuite

First, the NCBI NF1 mRNA sequence (NM_000267) was input into El Dorado to check for existing annotation and the PromoterInspector prediction in this region (Table 4, Figure 18). The program links this cDNA to chromosome 17 locus GX_025981, which is between position 4161289 – 4450606 (289318 bp) in sequence NT_010799. The output for this region includes annotated features like the 5' UTR, exons, introns, and repeat regions. Part of these results were summarized using Sockeye (Figure 20). However, the version of the sequence used in El Dorado and consequently the locations of the features are outdated. A BLAT search of this segment in UCSC showed the more updated location of the segment to be Hchr17:29267015-29556374, which is 289360 bp long instead of 289318 bp. Because my interest was on the PromoterInspector predictions, the locations of these predictions were determined in the current version of the human genome using BLAT in UCSC.

According to this annotation, the promoter region for NF1 gene is found at Hchr17:29271515-29272115 (601 bp). This region is –771 bp to –111 bp upstream of the translation start site. It does not include the ORF but spans the major TSS (Figure 21). No TATA box or other core promoter element was found. Two other promoters are predicted on the positive strand. One was found at Hchr17:29428959-29429559, which was a predicted gene AK024873 that spans NF1 intron 23.2, exon 23a, and intron 23a. The other is found at Hchr17:29527130-29522330 for a predicted gene AK025926 within NF1 intron 47. Both of these predicted genes are based on their mRNAs in the NCBI mRNA library. Three other promoters at Hchr17:29474355-

Table 4. Summary of El Dorado output of the human *NF1* gene. All positions are relative to position 4161289 in sequence NT_010799.

Pos. from	Pos. to	Strand	Length	Type of element	Name	Annotation
4501	5101	+	601 bp	Gene-associated Promoter	Associated with NF1 (NM_000267) Quality=silver (predicted promoter)	Predicted by PromoterInspector, 100 bp overlap with 5' end of NF1. Overlap with 5'UTR of NF1.
4701	5480	n/a	780 bp	PromoterInspect or Prediction	PI017010	479 bp overlap with 5' end of NF1. Overlap with 5'UTR of NF1.
5001	284318	+	279318 bp	Primary Transcript	NF1 (NM_000267)	Neurofibromin 1 (neurofibromatosis, von Recklinghausen disease, Watson disease)
5001	5271	+	271 bp	Exon	NF1 (NM_000267)	Exon 1
5001	5211	+	211 bp	5'UTR	NF1 (NM_000267)	-
5272	65886	+	60615 bp	Intron	NF1 (NM_000267)	Intron 1
65887	66030	+	144 bp	Exon	NF1 (NM_000267)	Exon 2
66031	68913	+	2883 bp	Intron	NF1 (NM_000267)	Intron 2
68914	68997	+	84 bp	Exon	NF1 (NM_000267)	Exon 3
68998	73089	+	4092 bp	Intron	NF1 (NM_000267)	Intron 3
73090	73280	+	191 bp	Exon	NF1 (NM_000267)	Exon 4
73281	79794	+	6514 bp	Intron	NF1 (NM_000267)	Intron 4
79795	79901	+	107 bp	Exon	NF1 (NM_000267)	Exon 5
79902	91325	+	11424 bp	Intron	NF1 (NM_000267)	Intron 5
91326	91393	+	68 bp	Exon	NF1 (NM_000267)	Exon 6
91394	91613	+	220 bp	Intron	NF1 (NM_000267)	Intron 6
91614	91689	+	76 bp	Exon	NF1 (NM_000267)	Exon 7
91690	92411	+	722 bp	Intron	NF1 (NM_000267)	Intron 7
92412	92569	+	158 bp	Exon	NF1 (NM_000267)	Exon 8
92570	110325	+	17756 bp	Intron	NF1 (NM_000267)	Intron 8
110326	110499	+	174 bp	Exon	NF1 (NM_000267)	Exon 9
110500	110940	+	441 bp	Intron	NF1 (NM_000267)	Intron 9
110941	111063	+	123 bp	Exon	NF1 (NM_000267)	Exon 10
111064	111314	+	251 bp	Intron	NF1 (NM_000267)	Intron 10
111315	111389	+	75 bp	Exon	NF1 (NM_000267)	Exon 11
111390	116143	+	4754 bp	Intron	NF1 (NM_000267)	Intron 11
116144	116275	+	132 bp	Exon	NF1 (NM_000267)	Exon 12
116276	124354	+	8079 bp	Intron	NF1 (NM_000267)	Intron 12
124355	124489	+	135 bp	Exon	NF1 (NM_000267)	Exon 13
124490	128908	+	4419 bp	Intron	NF1 (NM_000267)	Intron 13
128909	129022	+	114 bp	Exon	NF1 (NM_000267)	Exon 14
129023	131753	+	2731 bp	Intron	NF1 (NM_000267)	Intron 14
131754	131833	+	80 bp	Exon	NF1 (NM_000267)	Exon 15
131834	133347	+	1514 bp	Intron	NF1 (NM_000267)	Intron 15
133348	133471	+	124 bp	Exon	NF1 (NM_000267)	Exon 16
133472	134998	+	1527 bp	Intron	NF1 (NM_000267)	Intron 16
134999	135154	+	156 bp	Exon	NF1 (NM_000267)	Exon 17
135155	136338	+	1184 bp	Intron	NF1 (NM_000267)	Intron 17
136339	136588	+	250 bp	Exon	NF1 (NM_000267)	Exon 18
136589	137121	+	533 bp	Intron	NF1 (NM_000267)	Intron 18
137122	137195	+	74 bp	Exon	NF1 (NM_000267)	Exon 19
137196	137426	+	231 bp	Intron	NF1 (NM_000267)	Intron 19
137427	137510	+	84 bp	Exon	NF1 (NM_000267)	Exon 20
137511	138928	+	1418 bp	Intron	NF1 (NM_000267)	Intron 20
138929	139369	+	441 bp	Exon	NF1 (NM_000267)	Exon 21
139370	139738	+	369 bp	Intron	NF1 (NM_000267)	Intron 21
139739	139878	+	140 bp	Exon	NF1 (NM_000267)	Exon 22
139879	140163	+	285 bp	Intron	NF1 (NM_000267)	Intron 22
140164	140286	+	123 bp	Exon	NF1 (NM_000267)	Exon 23
140287	140745	+	459 bp	Intron	NF1 (NM_000267)	Intron 23
140746	140829	+	84 bp	Exon	NF1 (NM_000267)	Exon 24
140830	141977	+	1148 bp	Intron	NF1 (NM_000267)	Intron 24
141978	142094	+	117 bp	Exon	NF1 (NM_000267)	Exon 25
142095	142604	+	510 bp	Intron	NF1 (NM_000267)	Intron 25
142605	142786	+	182 bp	Exon	NF1 (NM_000267)	Exon 26
142787	142906	+	120 bp	Intron	NF1 (NM_000267)	Intron 26
142907	143118	+	212 bp	Exon	NF1 (NM_000267)	Exon 27
143119	145515	+	2397 bp	Intron	NF1 (NM_000267)	Intron 27
145516	145677	+	162 bp	Exon	NF1 (NM_000267)	Exon 28
145678	145822	+	145 bp	Intron	NF1 (NM_000267)	Intron 28

145823	145926	+	104 bp	Exon	NF1 (NM_000267)	Exon 29
145927	158888	+	12962 bp	Intron	NF1 (NM_000267)	Intron 29
158889	159024	+	136 bp	Exon	NF1 (NM_000267)	Exon 30
159025	168248	+	9224 bp	Intron	NF1 (NM_000267)	Intron 30
161948	162548	+	601 bp	Gene-associated Promoter	Associated with AK024873 (AK024873) Quality=gold (experimentally verified promoter)	Based on oligo_capped AK024873, 100 bp overlap with 5' end of AK024873. Overlap with exon 1 of AK024873. Within intron 30 of NF1.
162448	164178	+	1731 bp	Primary Transcript	AK024873 (AK024873)	Full length cDNA based on oligo capping method.
162448	164178	+	1731 bp	Exon	AK024873 (AK024873)	Exon 1
162448	162448	+	1 bp	Transcription Start Site	TSS	Oligo_capped AK024873_1
168249	168407	+	159 bp	Exon	NF1 (NM_000267)	Exon 31
168408	168936	+	529 bp	Intron	NF1 (NM_000267)	Intron 31
168937	169034	+	98 bp	Exon	NF1 (NM_000267)	Exon 32
169035	170273	+	1239 bp	Intron	NF1 (NM_000267)	Intron 32
170274	170420	+	147 bp	Exon	NF1 (NM_000267)	Exon 33
170421	171615	+	1195 bp	Intron	NF1 (NM_000267)	Intron 33
171616	171762	+	147 bp	Exon	NF1 (NM_000267)	Exon 34
171763	175128	+	3366 bp	Intron	NF1 (NM_000267)	Intron 34
175129	175239	+	111 bp	Exon	NF1 (NM_000267)	Exon 35
175240	235686	+	60447 bp	Intron	NF1 (NM_000267)	Intron 35
204461	207423	-	2963 bp	Primary Transcript	OMG (NM_002544)	Oligodendrocyte myelin glycoprotein (PubMed) more gene info...
204461	207423	-	2963 bp	Exon	OMG (NM_002544)	Exon 1
204461	204895	-	435 bp	3'UTR	OMG (NM_002544)	
207323	207923	-	601 bp	Gene-associated Promoter	Associated with OMG (NM_002544) Quality=bronze (5' upstream region)	100 bp overlap with 5' end of OMG. Overlap with exon 1 of OMG. Within intron 35 of NF1.
213658	223947	-	10290 bp	Primary Transcript	EVI2B (NM_006495)	Ecotropic viral integration site 2B (PubMed) more gene info...
213658	215514	-	1857 bp	Exon	EVI2B (NM_006495)	Exon 2
213658	214145	-	488 bp	3'UTR	EVI2B (NM_006495)	
215494	215514	-	21 bp	5'UTR	EVI2B (NM_006495)	
215515	223852	-	8338 bp	Intron	EVI2B (NM_006495)	Intron 1
223847	224447	-	601 bp	Gene-associated Promoter	Associated with EVI2B (NM_006495) Quality=bronze (5' upstream region)	100 bp overlap with 5' end of EVI2B. Within intron 35 of NF1. Overlap with 5'UTR of EVI2B.
223853	223947	-	95 bp	Exon	EVI2B (NM_006495)	Exon 1
223853	223947	-	95 bp	5'UTR	EVI2B (NM_006495)	
227533	231564	-	4032 bp	Primary Transcript	EVI2A (NM_014210)	Ecotropic viral integration site 2A (PubMed) more gene info...
227533	228898	-	1366 bp	Exon	EVI2A (NM_014210)	Exon 2
227533	228177	-	645 bp	3'UTR	EVI2A (NM_014210)	
228877	228898	-	22 bp	5'UTR	EVI2A (NM_014210)	
228899	231367	-	2469 bp	Intron	EVI2A (NM_014210)	Intron 1
231368	231564	-	197 bp	Exon	EVI2A (NM_014210)	Exon 1
231368	231564	-	197 bp	5'UTR	EVI2A (NM_014210)	
231464	232064	-	601 bp	Gene-associated Promoter	Associated with EVI2A (NM_014210) Quality=bronze (5' upstream region)	100 bp overlap with 5' end of EVI2A. Within intron 35 of NF1. Overlap with 5'UTR of EVI2A.
235687	236119	+	433 bp	Exon	NF1 (NM_000267)	Exon 36
236120	237363	+	1244 bp	Intron	NF1 (NM_000267)	Intron 36
237364	237704	+	341 bp	Exon	NF1 (NM_000267)	Exon 37
237705	240413	+	2709 bp	Intron	NF1 (NM_000267)	Intron 37
240414	240616	+	203 bp	Exon	NF1 (NM_000267)	Exon 38
240617	244589	+	3973 bp	Intron	NF1 (NM_000267)	Intron 38
244590	244783	+	194 bp	Exon	NF1 (NM_000267)	Exon 39
244784	246191	+	1408 bp	Intron	NF1 (NM_000267)	Intron 39
246192	246332	+	141 bp	Exon	NF1 (NM_000267)	Exon 40
246333	246494	+	162 bp	Intron	NF1 (NM_000267)	Intron 40
246495	246774	+	280 bp	Exon	NF1 (NM_000267)	Exon 41
246775	247227	+	453 bp	Intron	NF1 (NM_000267)	Intron 41
247228	247442	+	215 bp	Exon	NF1 (NM_000267)	Exon 42
247443	247679	+	237 bp	Intron	NF1 (NM_000267)	Intron 42
247680	247741	+	62 bp	Exon	NF1 (NM_000267)	Exon 43
247742	247885	+	144 bp	Intron	NF1 (NM_000267)	Intron 43
247886	248000	+	115 bp	Exon	NF1 (NM_000267)	Exon 44
248001	248565	+	565 bp	Intron	NF1 (NM_000267)	Intron 44

248566	248667	+	102 bp	Exon	NF1 (NM_000267)	Exon 45
248668	250366	+	1699 bp	Intron	NF1 (NM_000267)	Intron 45
250367	250507	+	141 bp	Exon	NF1 (NM_000267)	Exon 46
250508	252884	+	2377 bp	Intron	NF1 (NM_000267)	Intron 46
252885	253011	+	127 bp	Exon	NF1 (NM_000267)	Exon 47
253012	258998	+	5987 bp	Intron	NF1 (NM_000267)	Intron 47
254692	255292	+	601 bp	Gene-associated Promoter	Associated with AK025926 (AK025926) Quality=gold (experimentally verified promoter)	Based on oligo_capped AK025926, 100 bp overlap with 5' end of AK025926. Within intron 47 of NF1. Overlap with exon 1 of AK025926.
255173	255472	n/a	300 bp	PromoterInspector Prediction	PI017011	280 bp overlap with 5' end of AK025926. Within intron 47 of NF1. Overlap with exon 1 of AK025926.
255192	256903	+	1712 bp	Primary Transcript	AK025926 (AK025926)	Full length cDNA based on oligo capping method.
255192	256903	+	1712 bp	Exon	AK025926 (AK025926)	Exon 1
255192	255192	+	1 bp	Transcription Start Site	TSS	Oligo_capped AK025926_2
258999	259130	+	132 bp	Exon	NF1 (NM_000267)	Exon 48
259131	260061	+	931 bp	Intron	NF1 (NM_000267)	Intron 48
260062	260197	+	136 bp	Exon	NF1 (NM_000267)	Exon 49
260198	262134	+	1937 bp	Intron	NF1 (NM_000267)	Intron 49
262135	262292	+	158 bp	Exon	NF1 (NM_000267)	Exon 50
262293	266337	+	4045 bp	Intron	NF1 (NM_000267)	Intron 50
266338	266460	+	123 bp	Exon	NF1 (NM_000267)	Exon 51
266461	266837	+	377 bp	Intron	NF1 (NM_000267)	Intron 51
266838	266968	+	131 bp	Exon	NF1 (NM_000267)	Exon 52
266969	267146	+	178 bp	Intron	NF1 (NM_000267)	Intron 52
267147	267247	+	101 bp	Exon	NF1 (NM_000267)	Exon 53
267248	268357	+	1110 bp	Intron	NF1 (NM_000267)	Intron 53
268358	268500	+	143 bp	Exon	NF1 (NM_000267)	Exon 54
268501	268846	+	346 bp	Intron	NF1 (NM_000267)	Intron 54
268847	268893	+	47 bp	Exon	NF1 (NM_000267)	Exon 55
268894	270364	+	1471 bp	Intron	NF1 (NM_000267)	Intron 55
270365	270581	+	217 bp	Exon	NF1 (NM_000267)	Exon 56
270582	283884	+	13303 bp	Intron	NF1 (NM_000267)	Intron 56
283885	284318	+	434 bp	Exon	NF1 (NM_000267)	Exon 57
284028	284318	+	291 bp	3'UTR	NF1 (NM_000267)	
284681	285281	+	601 bp	Gene-associated Promoter	Associated with AK026658 (AK026658) Quality=gold (experimentally verified promoter)	Based on oligo_capped AK026658, 100 bp overlap with 5' end of AK026658. Overlap with exon 1 of AK026658.
285181	287541	+	2361 bp	Primary Transcript	AK026658 (AK026658)	Full length cDNA based on oligo capping method.
285181	287541	+	2361 bp	Exon	AK026658 (AK026658)	Exon 1
285181	285181	+	1 bp	Transcription Start Site	TSS 1	Oligo_capped AK026658_

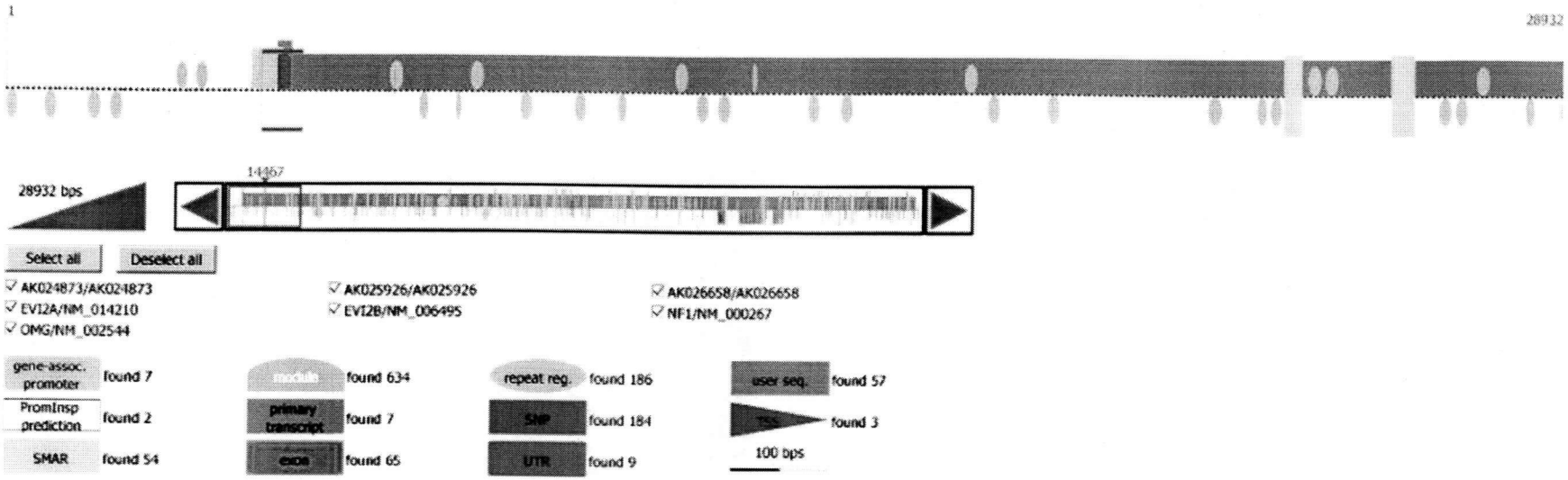


Figure 18. Output of El Dorado for human *NF1*. Features shown include the gene-associated promoter, PromoterInspector Prediction, primary transcript, exons, and UTRs.

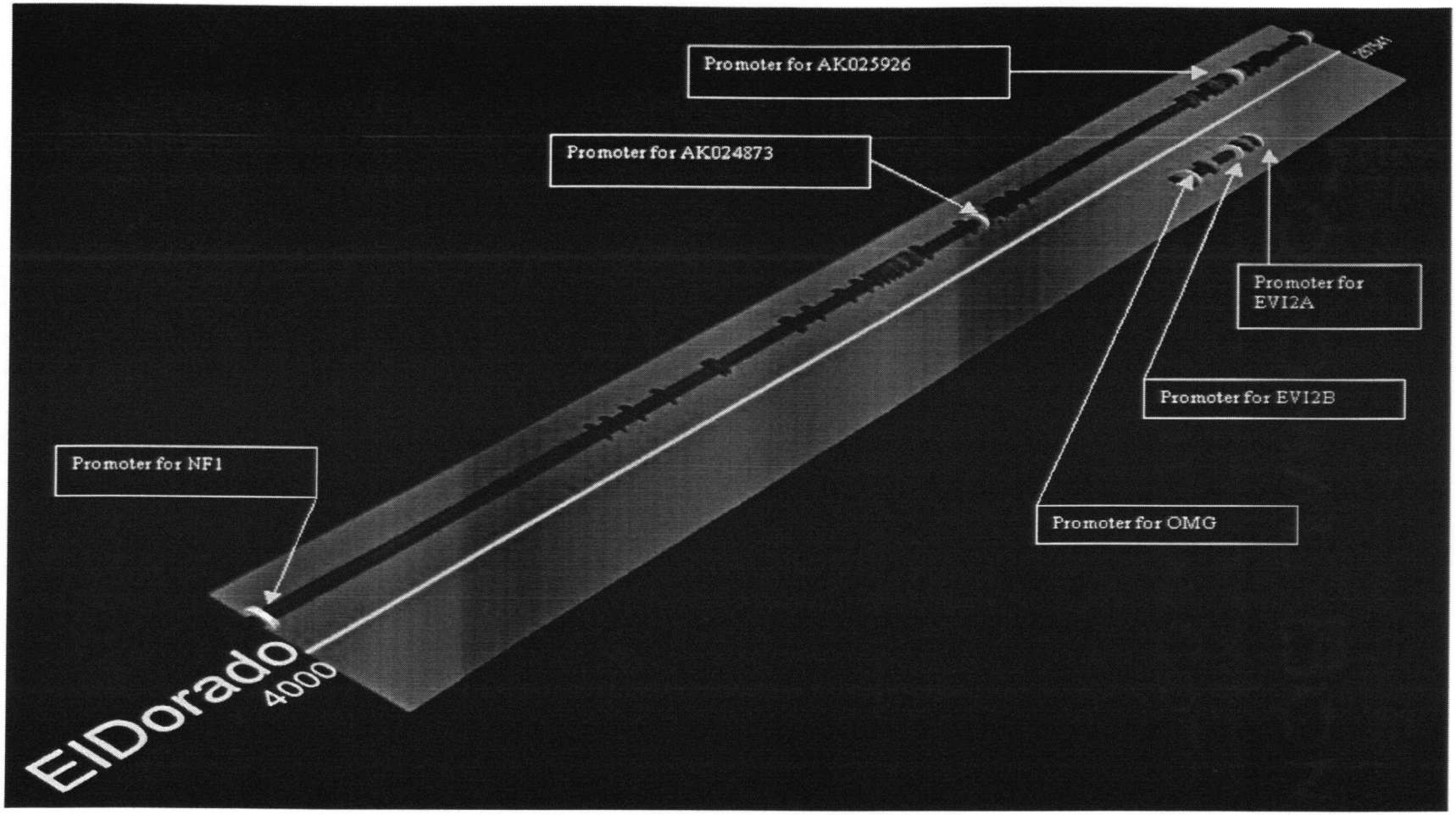


Figure 19. Graphical presentation of El Dorado output for human *NF1* gene using Sockeye. Introns are shown as blue rods, and exons are shown as purple rods. Predicted promoter regions are shown in yellow semicircles.

29474955, 29490889-29491489, and 29498507-29499105 are identified for the genes OMV, EVI2B, and EVI2A coded on the reverse strand (Figure 19).

3.2.2 Dragon Promoter Finder

With the restriction of sequence length and the TSS (484 bp upstream of ORF) position in mind, a test sequence was constructed to include at least 684 bp upstream of the translation start site (200 bp upstream of TSS) and a length of 10,000 bp. Hchr17:29271526-29281525 is 10,000 bp long and includes 700 bp upstream of the translation start site (216 bp upstream of the TSS). Sensitivity settings of 80%, 65%, and 50% yielded results that were upstream of the ORF and on the positive strand. 35% and 25% (higher specificity) did not.

Most promoter regions are found between -250 bp (for TATA box and proximal promoter regions) and +32 bp (for DPE) of the TSS (Kadonaga, 2002). At 80% sensitivity, DPF predicts 8 potential TSSs, only two of which were upstream of the ORF at 384 and 116 bp upstream of the translation start site. Therefore, after extending 250 bp upstream and 32 bp downstream of the predicted TSSs, the two predicted promoter regions would be –534 bp to –353 bp (Hchr17:29271692-29271873) and –366 bp to –85 bp (Hchr17:29271860-29272141) relative to the translation start site (+1). All the other predictions were downstream of the ORF (+127, +637, +551, +3352, +4043, +5600), so they are unlikely to represent the *NF1* TSS. At 60% sensitivity, three predictions were made, and the only TSS detected upstream of the ORF was 383 bp upstream. At 50% sensitivity, only the TSS 116 bp upstream of the ORF start site was detected. These results are summarized with the PromoterInspector result in Figure 20.

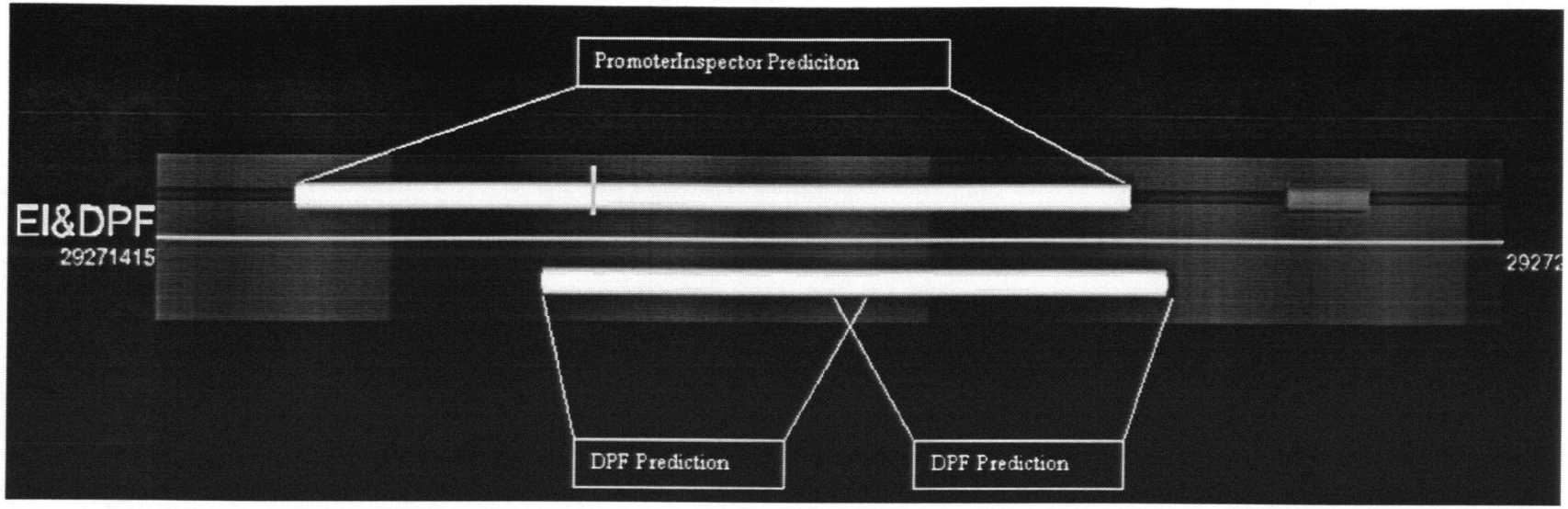


Figure 20. Comparison of El Dorado (PromoterInspector) and Dragon Promoter Finder Predictions for the potential promoter region of human *NF1* gene. Introns and 5UR are shown as blue rods. Exon 1 is shown as a purple rod. The yellow rods are the predicted promoter regions. The grey line is the TSS. DPF predictions are located on the positive strand, but they are shown below the main strand for clarity. Note that the two DPF predictions (Hchr17:29271692-29271873 and Hchr17:29271860-29272141) overlap.

3.3 Comparison of the Human, Mouse, Rat, and Pufferfish *NF1* ORF and Protein

mRNAs of human (H), rat (R), pufferfish (F) were obtained from NCBI with the following accession numbers NM_000267, NM_012609, and AF064564, while mRNA of *Fugu* was obtained from ENSEMBL with accession number ENSMUST00000000890. The 5' UTR and 3' UTR of each sequence was trimmed to form the ORF, which includes the translation start site, all the normally-spliced exons, and the alternatively spliced exon 23b. Note that *Fugu* does not have exon 12b. These sequences were then compared with mVista and have the identities shown in Table 5.

These ORFs were translated into amino acid sequences and are compared with Pairwise BLAST under default settings using blastp. The results are shown in Table 5.

3.4 Defining the 5' Upstream Region (5UR)

Because this study is designed to search for potential tfactor binding sites, the intronic regions, and especially the region upstream of the *NF1* ORF, were of special interest. Although it is possible that regions that regulate transcription lie upstream of the 5' flanking gene of NF1, a valid tfactor binding site within or upstream of the 5' flanking gene is much more likely to be relevant to transcription of that gene than to *NF1*. Therefore, for the purpose of this study, the 5' upstream region (5UR) was defined as the intragenic region between the *NF1* translation start site and the 3' end of the coding region of the gene that is immediately 5' of *NF1* in human.

Table 5. Identities of *NF1* ORF genomic nucleotide sequence and protein sequence among human, mouse, rat, and pufferfish. Green boxes represent percent identities of nucleotides within the *NF1* ORF. Yellow boxes represent protein identities.

	Human	Mouse	Rat	Pufferfish
Human		97.00%	95.35%	83.40%
Mouse	91.45%		95.74%	84.24%
Rat	89.42%	93.46%		82.39%
Pufferfish	70.27%	71.23%	69.74%	

According to UCSC, the first annotated gene 5' of the *NF1* translation start site (Hchr17:29272226) is NT_010799.114, a GenScan prediction at Hchr17:29148001-29212469 (Figure 21). This gene has a supportive mRNA (AF086476) and spliced ESTs (AV764031, W84334, AI1683491, and BG182434). On the other hand, MGC13061 at Hchr17:29147950-29176825 is the closest known functional gene upstream of the human *NF1* gene that has the same transcription direction to date (Jenne *et al.*, 2003). It is over 95000 nucleotides away from the *NF1* translation start site while NT_010799.144 is less than 60000 nucleotides away. Furthermore, NT_010799.114 overlaps with MGC13061, so it could be a separate gene or part of MFC13061, if the annotation of MGC13061 does not reflect its entire length. There is no known or predicted gene annotated closer to the *NF1* gene, so NT_010799.114 was used as the 5' flanking gene for NF1. The NF1 5UR was, therefore, defined as Hchr17:29212470-29272225 (59756 bp).

Unlike the human *NF1* gene, the mouse *NF1* gene is not yet annotated, but it is represented as gene AK085050 in UCSC. Using ENSEMBL data on mouse *NF1* and BLAT in UCSC, mouse *NF1* exon 1 was found to begin at Mchr11:80125292, which is also the translation start site. According to UCSC, the first annotated gene upstream of the *NF1* gene on the same strand is *Nos2*, which was located 419149 bp away at Mchr11:79706143-79745518 (Figure 22). Using mouse Nos2 mRNA in BLAT, this gene is located at chr17:25935666-25979387 on the reverse strand in human, and at chr10:61267669-61303238 on the forward strand in rat. If 5UR in mouse is defined to be between the *NF1* gene and the *Nos* gene, the size of the *NF1* 5UR is 379773 bp, which is exceedingly large. For reasonable alignment and comparison with human 5UR, a relatively similar length is needed. Therefore, the mouse 5UR was defined to have the same length as human 5UR and to extend from Mchr11:80065536-80125291.

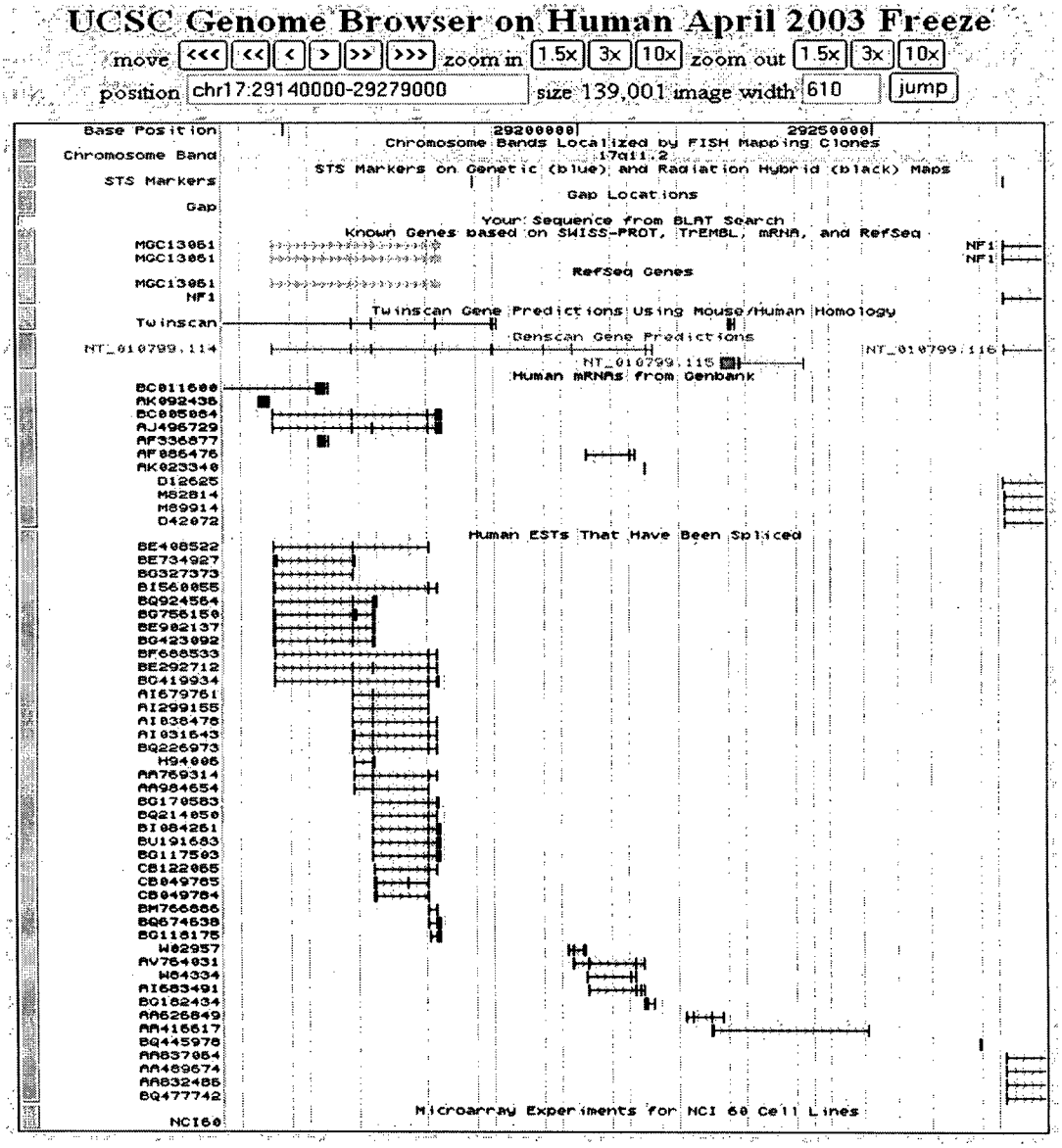


Figure 21. UCSC annotation in region Hchr17:29140000-29279000. Note the position of the *NF1* gene near the upper right corner of the figure and the closest annotated upstream gene MGC13061. Also note the position of the GenScan prediction NT010779.114 and its supporting EST data.

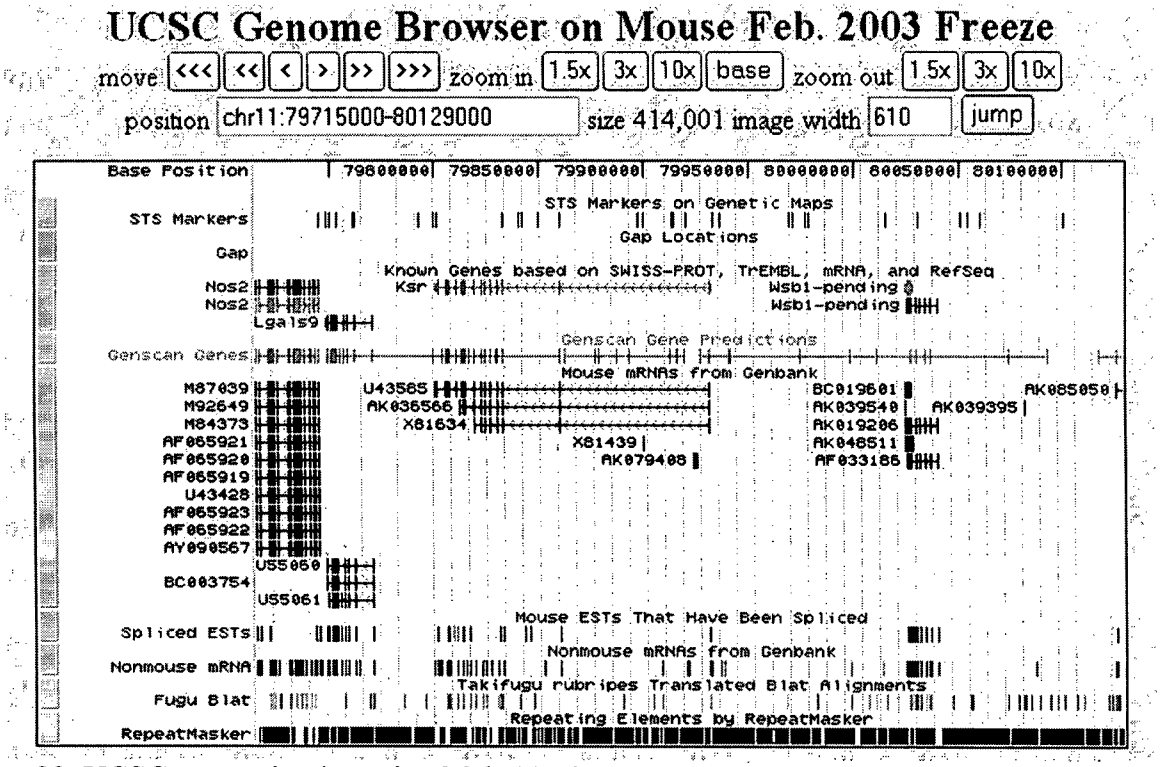


Figure 22. UCSC annotation in region Mchr11:79715000-80129000. Note the position of the *NF1* gene (AK085050) at the right edge of the figure and the position of the annotated gene *Nos2*. Also note the supportive mRNA and ESTs data for the *Nos* gene. Note that both *Krs* amd *Wsb1-pending* are coded on the opposite strand.

Although the rat *NF1* gene is annotated in UCSC, the region around the translation start site is not completely sequenced. Because mouse and rat share high homology, mouse intron 1 (Mchr11:80125352-80150351) was used to estimate the position of the rat translation start site. BLAT indicates that Rchr10:61761033 corresponds to Mchr11:80125357. Assuming the distance between the mouse translation start site (Mchr11: 80125292) and Mchr11:80125357 is identical to the distance between rat translation start site and Rchr10:61761033, then the rat translation start site will be Rchr10:61760968. The first annotated gene in rat upstream of the *NF1* gene on the same strand is also *Nos2*, which is located 133364 bp away at Rchr10:61267669-61303238 (Figure 23). Again, defining the rat 5UR based on the distance between the *NF1* gene and *Nos2* gene would be too large for this study, so the 5UR in rat was defined as Rchr10:61701212-61760967 , i.e., of the same length as human and mouse 5UR.

The *Fugu* genome is sequenced but not assembled, so the *Fugu NF1* gene is represented in NCBI by a series of contigs. Annotation is not available, but the 5' flanking gene of *Fugu NF1* has been described (Kehrer-Sawatzki. 2000). The gene is called FN5 (flanking *Fugu* NF1 gene in 5' direction). Using Pairwise BLAST, the sequence reported as FN5 ends at nucleotide 21241 of contig CAAB01003481. The same result is confirmed by GenScan. Also, using exon data of *Fugu* supplied by NCBI, nucleotide 22730 of contig CAAB01003481 corresponds to the translational start site of FN5. Therefore, the *Fugu NF1* 5UR region was defined as CAAB01003481:21242-22729. Note that this region is only 1488 bp long, compared to the nearly 6 kb long 5UR of human.

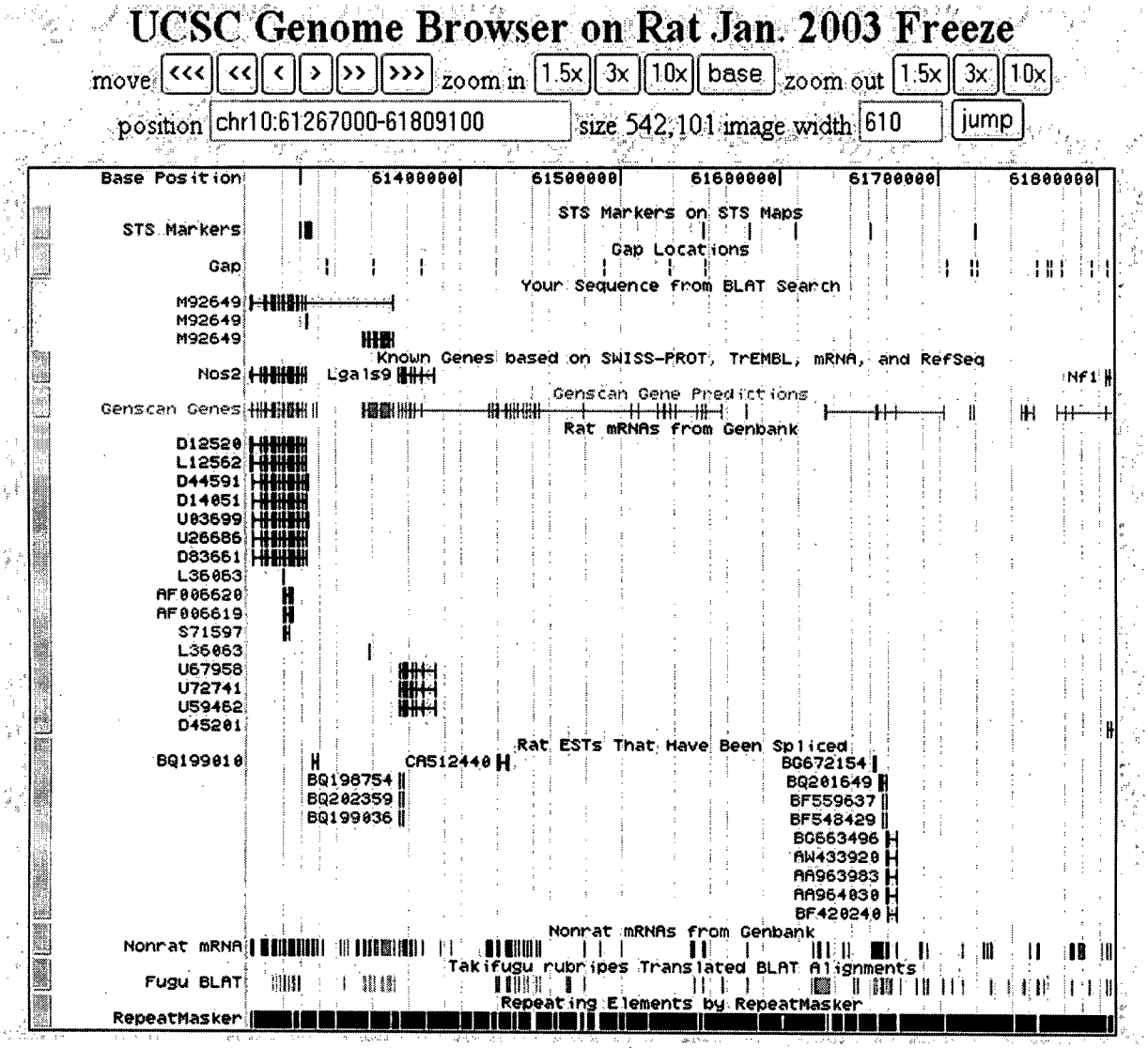


Figure 23. UCSC annotation in region Rchr10:61267000-61809100. Note the position of the *NF1* gene on the right edge of the figure and the closest annotated gene *Nos2*, which has supportive EST data. Note that *Lgal1s9* is coded on the opposite strand.

3.5 Defining Exon-Intron 1 (EI1)

For the purpose of this study, exon 1 is defined as the coding region of mRNA before exon 2. The 5' UTR is not included in this definition. For human, mouse, rat and pufferfish, this definition will lead to a common length of 60 bp for exon 1 instead of a different length in each species if the 5' UTR is included. Exon-Intron 1 (EI1) is defined as the combination of exon 1 and intron 1. The human, mouse, and rat EI1 regions were obtained from UCSC as Hchr17:29272226-29332898, Mchr11:80125292-80169466, and Rchr10:61760968-61804379, respectively. The pufferfish EI1 region is covered by contig CAAB01003481: 22730-25367. The lengths of EI1 in human, mouse, rat, and pufferfish are 60673, 44175, 43412, and 2638 bp, respectively.

3.6 Comparison of *NF1* 5UR and EI1 in Human (H), Mouse (M), Rat (R), and Pufferfish (F)

Comparison of the *NF1* 5UR and EI1 regions were performed between human, mouse, rat and pufferfish. Each comparison was divided into two sections, and the first section was divided into two parts. Section 1a was a search for 'windows' of high homology shared by human, mouse and rat. The location of these windows was pinpointed by using mVISTA for sequence alignment, followed by analysis with Frameslider.

The cut-off for window selection was based on the amount of homology observed in the *NF1* ORF for each pair of species (Table 5). The cutoff values used to define "high homology" in non-coding regions were HvsM – 0.90, HvsR – 0.88, HvsF – 0.70, MvsR – 0.92, MvsF – 0.70,

and RvsF – 0.68. These cut-off values are not exactly the same as the percent identities of the ORFs because with a window size of 50, identity jumps in increments of 0.02. For example, the observed HvsM identity in the ORF was 91.45% (Table 5). Either 0.90 or 0.92 could be chosen as the cutoff, but the higher value may be so high that crucial homologies are lost. Therefore, a cut-off value of 0.90 was chosen. Similarly, for other cut-off values, the closest value smaller than the corresponding ORF identity was used. The rationale behind these cut-off values is that if a non-coding region shares as high an identity as the functional coding sequences when compared to other species, then it is very likely to be functional. Adjacent or overlapping windows were combined to form a bigger region when shared by human, mouse, and rat. In this case, the identity of a large homologous region is reported as a range if there are differences among the identities of the individual windows. If a highly homologous region appeared likely not to be functional upon manual inspection due to potential repeat elements, RepeatMasker was used to confirm this observation.

Section 1b of each analysis focused on comparison between the *Fugu* sequence and the highly homologous non-coding regions observed in mammals. Note that analyses related to *Fugu* were not included in the definition of the highly homologous regions in Section 1a for several reasons. First, because of the extreme difference in length between *Fugu* and other organisms, only the human, mouse, and rat non-coding sequences could be aligned along their full lengths using mVista. Furthermore, because of the extreme difference in evolutionary distance, there is a low possibility that sequence would be conserved over a large window size. Therefore, each highly homologous region found in mammals was compared to the *Fugu* sequence using Pairwise BLAST with two different settings. One was the default blastn setting, except the "Forward strand" option was chosen instead of "Both strands". The other blastn setting used was the default except the Penalty for a mismatch was set at -20, the Open gap penalty was set at -20, the

Extended gap penalty was set at -20, the Expect was set at – 99999999, and the Filter was set at – Off. Wordsize was iteratively increased from 7 (its minimum value) until the lowest number that produced no hits was found. These matches between the human and *Fugu* sequences were then compared with the mouse and rat sequences to confirm the identity. This method was designed to search for nucleotides that have the longest exact match between *Fugu* and the mammalian species. Although there is no maximum value for 'Expect', the stringency setting was very low as 'Expect' was set to be extremely large.

Furthermore, since there is a higher probability that tfactor binding sites that influence *NF1* transcription are located closer to the beginning of the translation start site rather than thousands of bp away, an mVista analysis was done with the 1488 bp *Fugu* 5UR. In this alignment, the whole *Fugu* 5UR was aligned with the 1488 bps immediately upstream of translational start site of human, mouse, and rat *NF1*. Because of the extreme difference in evolutionary distance, Frameslider was used with the *Fugu* 5UR as the primary sequence and a window size of 20, instead of the window size of 50 used for comparisons among human, mouse, and rat sequences.

Section 2 of each analysis was the tfactor binding site prediction. For each region of high homolgy that was not due to repeat elements, the sequence including each highly homologous region and extending 100 bp upstream and downstream was analyzed using MATCH™ and MatInspector. For MATCH™, predictions were made using the minFN, minFP, and minSUM settings. For MatInspector, predictions were made switching the core similarity settings between 0.70 (lowest) and 1.00 (highest) while keeping the matrix similarity setting at 'optimized'. Predictions from these programs were then compared, with special attention paid to predictions that indicated a common tfactor at the same aligned position. All results were summarized using Sockeye.

3.7 Analyses of the 5UR

Using mVista and Frameslider, 3 regions of similar or greater homology than the *NF1* coding region were found in the comparison of the 5UR in human vs. mouse, 9 in the comparison of human vs. rat, and 144 in the comparison of mouse vs. rat. In this study, a highly homologous region was defined as a sequence of at least 50 bp shared by human, mouse, and rat. Therefore, homology regions obtained from the HvsM comparison, which yielded the smallest number of candidate regions, were searched for among the segment found to show homology as great or greater than that of the coding reiongs in comparisons of HvsR and MvsR. All three regions found in the HvsM comparison were also found in the HvsR and MvsR comparisons, so they were defined as highly homologous regions (Figure 24).

3.7.1 5UR-HHR1 (5' Upstream Region Highly Homologous Region 1) - Section 1a (H, M, & R)

5UR-HHR1 is located at Hchr17:29229534-29229600, which aligns to Mchr11:80092345-80092416 and Rchr10:61725619-61725686 (Figure 25). Frameslider indicates that identity for this segment between HvsM is 0.90, between HvsR is 0.88, and between MvsR is 0.82-0.90. However, inspection indicates that the alignment of mVista is incorrect in the MvsR comparison because unnecessary gaps are introduced at the end of the MvsR alignment. Once corrected, the identity ranges from 0.9-0.92. There is no repeat in this region according to RepeatMasker.

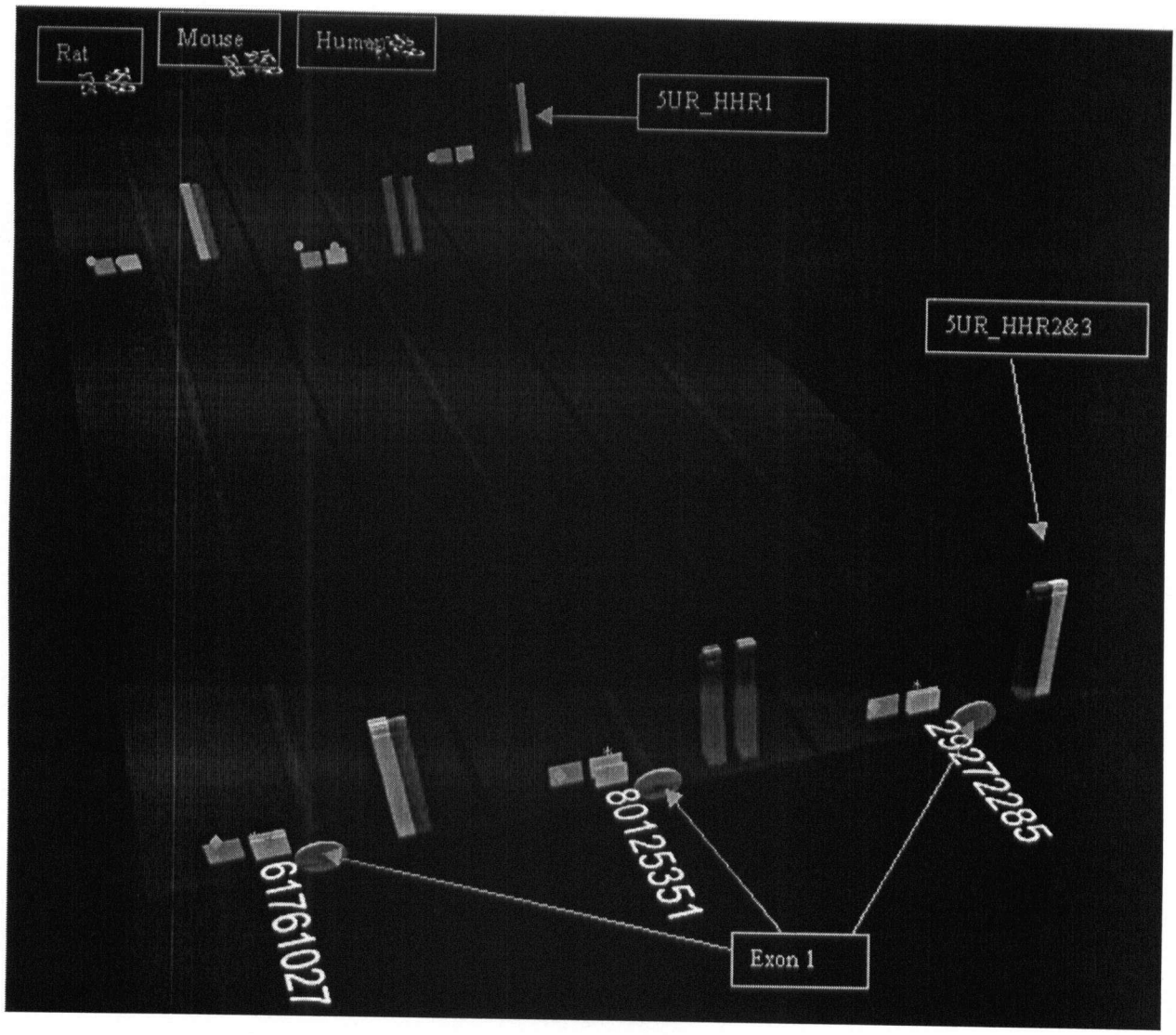


Figure 24. Summary for 5UR. Blue rods denote non-coding regions; purple circles denote exon 1. Bars perpendicular to the plane are highly homologous as indicated by the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. The bars also indicate the locations of the highly homologous regions shared by human, mouse, and rat. Tfactor predictions are indicated by boxes under the bar representing the 5UR in the follow colours: Yellow – MATCH^{TM}, Orange – MatInspector. Locations of some tfactor predictions are shown. Tfactor binding sites are represented by the following symbols: • for AP-1, ♦ for AP-2, * for Pax-4, and + for c-Ets-1 (p54).

a) mVista Alignment

```
HvsM
Human    gaacacgccctggaaacagaaaccgtct-gtcattc----taaggccattgtggttggga
         ||||| |||||||||||||||| | ||| |||| ||     ||||||||||||||| |||
Mouse    gaacaggccctggaaacagaaagcctcttgtcactcgctcgaaggccattgtggttagga

Human    gtgattcagact
         ||||||||||||
Mouse    gtgattcagact

HvsR
Human    gaacacgccctggaaacagaaaccgtct-gtcattctaaggccattgtggttgggagtga
         ||||| |||||||||||||||| ||||| |||  || ||||||||||||||| |||||||
Rat      gaacaggccctggaaacagaaagcctcttgtcgctcgaaggccattgtggttaggagtga

Human    ttcagact
         ||||||||
Rat      ttcagact

MvsR
Mouse    gaacaggccctggaaacagaaagcctcttgtcactcgctcgaaggccattgtggttagga
         |||||||||||||||||||||||||||||||| ||||    ||||||||||||||||||
Rat      gaacaggccctggaaacagaaagcctcttgtcgctcg----aaggccattgtggttagga

Mouse    gtgattcagact
         ||||||||||||
Rat      gtgattcagact
```

b) Combined Alignment

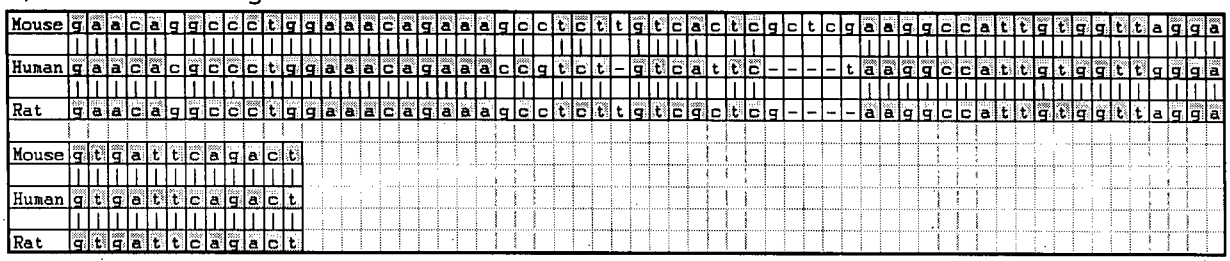


Figure 25. Alignment of 5UR-HHR1 on Hchr17:29229534-29229600, Mchr11:80092345-80092416 and Rchr10:61725619-61725686 from a) mVista and b) when combined. Nucleotides highlighted in yellow are shared by all three species. Note that the corrected alignment between mouse and rat is shown.

5UR-HHR1 falls within a GenScan predicted gene (NT_010799.115) located at Hchr17:29224077-29238353 on the reverse strand. This gene has supporting EST AA628349 (Figure 21, Figure 26). This EST does not overlap with 5UR-HHR1. To investigate the potential relationship of 5UR-HHR1 to this GenScan prediction, the following experiments were done. A region 1000 bp upstream and downstream of NT_010799.115 (chr17:29223077-29239353), which is 6457 upstream and 9753 downstream of the highly homologous region, was downloaded and analyzed with GenScan under its default settings to determine the structure of the predicted gene. The essential structures are shown in Table 6. Next, corresponding regions in the mouse and rat were inspected using UCSC. In this case, the regions were Mchr11:80085888-80102169 and Rchr10:61719162-61735438 (Figure 26). Although there were gene predictions for mouse (chr11_16.16) and rat (chr10_13.37), they were both upstream of 5UR-HHR1, with the prediction in mouse on the opposite strand and the prediction in rat on the same strand. Furthermore, Pairwise BLAST has shown that the predicted mRNAs of NT_010799.15, chr11_16.16, and chr10-13.37 were not homologous to one another. When the locations of the exons predicted by GenScan and the Frameslider results of HvsM, HvsR, and MvsR were all plotted on the same scale, 5UR-HHR1 was found not to be within an exon of the predicted gene (Figure 27). Sockeye shows the position of 5UR-HHR1 relative to the GenScan prediction for NT_010799.115 (Figure 28). Lastly, the mRNA of NT_010799.115 was downloaded and BLASTed against the human (default setting), mouse and rat (both at lower stringency with Expect at 10) EST databases in NCBI. 130 hits were found in the human database, (hit with the lowest expect value was 5' of *NF1*), but no hit was found in the mouse or rat database.

Therefore, the GenScan predicted gene NT_010799.115 may exist in humans but not in mice or rats. If this gene does exist in all 3 species, it is not in the same position relative to 5UR-HHR1

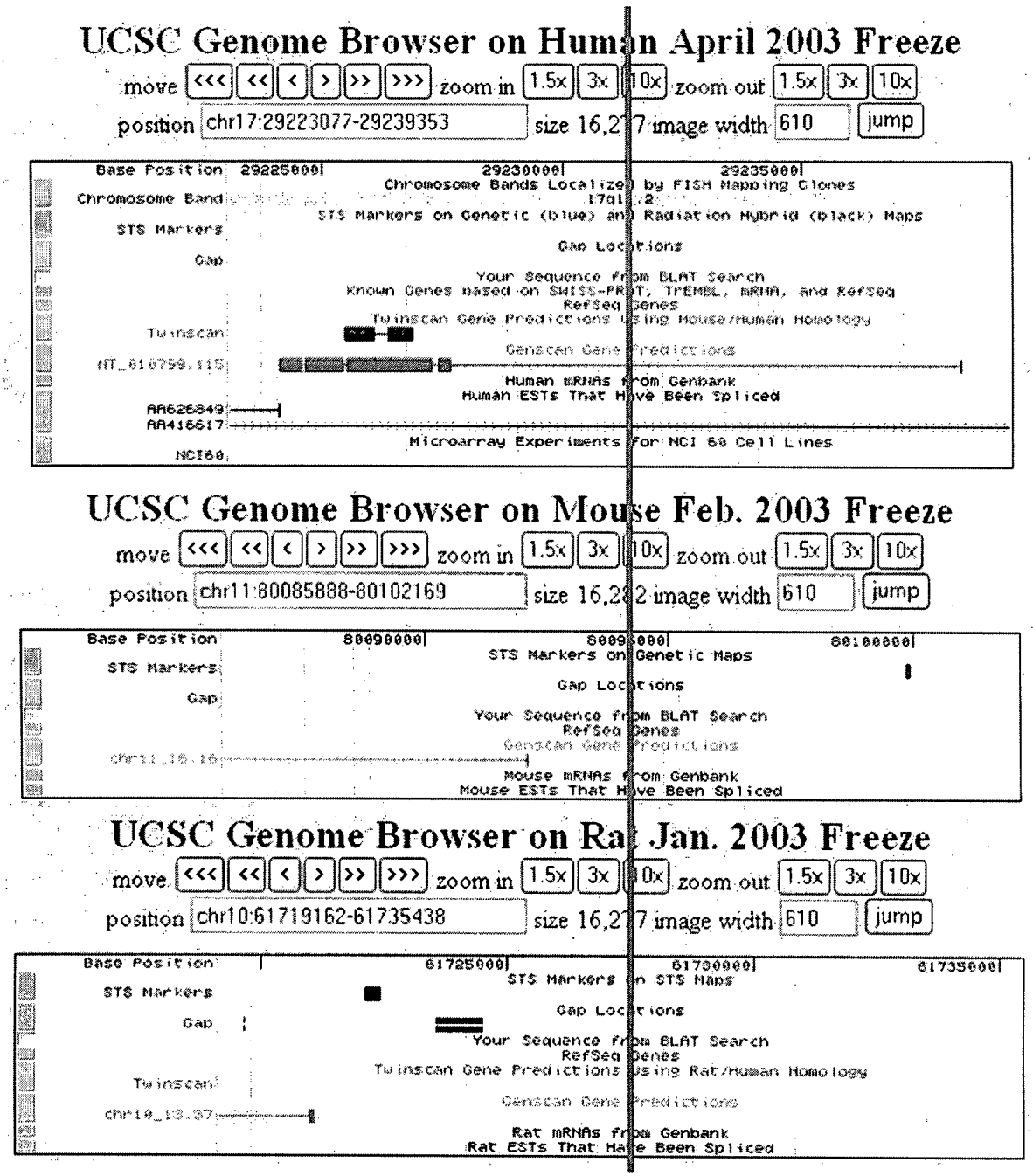


Figure 26. UCSC annotations at Hchr17:29223077-29239353, Mchr11:80085888-80102169, and Rchr10:61719162-61735438. The vertical green line denotes the location of 5UR-HHR1.

Table 6. GenScan prediction on NT_010799.115 in region chr17:29223077-29239353. The locations of exon features on chromosome 17 are shown. Predicted exons with higher probability values are more likely to be correct. Exon score > 100 is strong; 50-100 is moderate; 0-50 is weak; and below 0 is poor.

Gene number, exon number	Type	Strand	Beginning of exon/signal	End point of exon or signal	Length	Probability as exon	Exon Score
1.06	poly-A signal (consensus: AATAAA)	-	29223612	29223607	6		-0.45
1.05	Terminal exon (3' splice site to stop codon)	-	29224580	29224077	504	0.936	14.04
1.04	Internal exon (3' splice site to 5' splice site)	-	29225312	29224735	578	0.23	6.94
1.03	Internal exon (3' splice site to 5' splice site)	-	29226904	29225561	1344	0.302	39.63
1.02	Internal exon (3' splice site to 5' splice site)	-	29227676	29227466	211	0.959	19.59
1.01	Initial exon (ATG to 5' splice site)	-	29238353	29238276	78	0.253	1.66

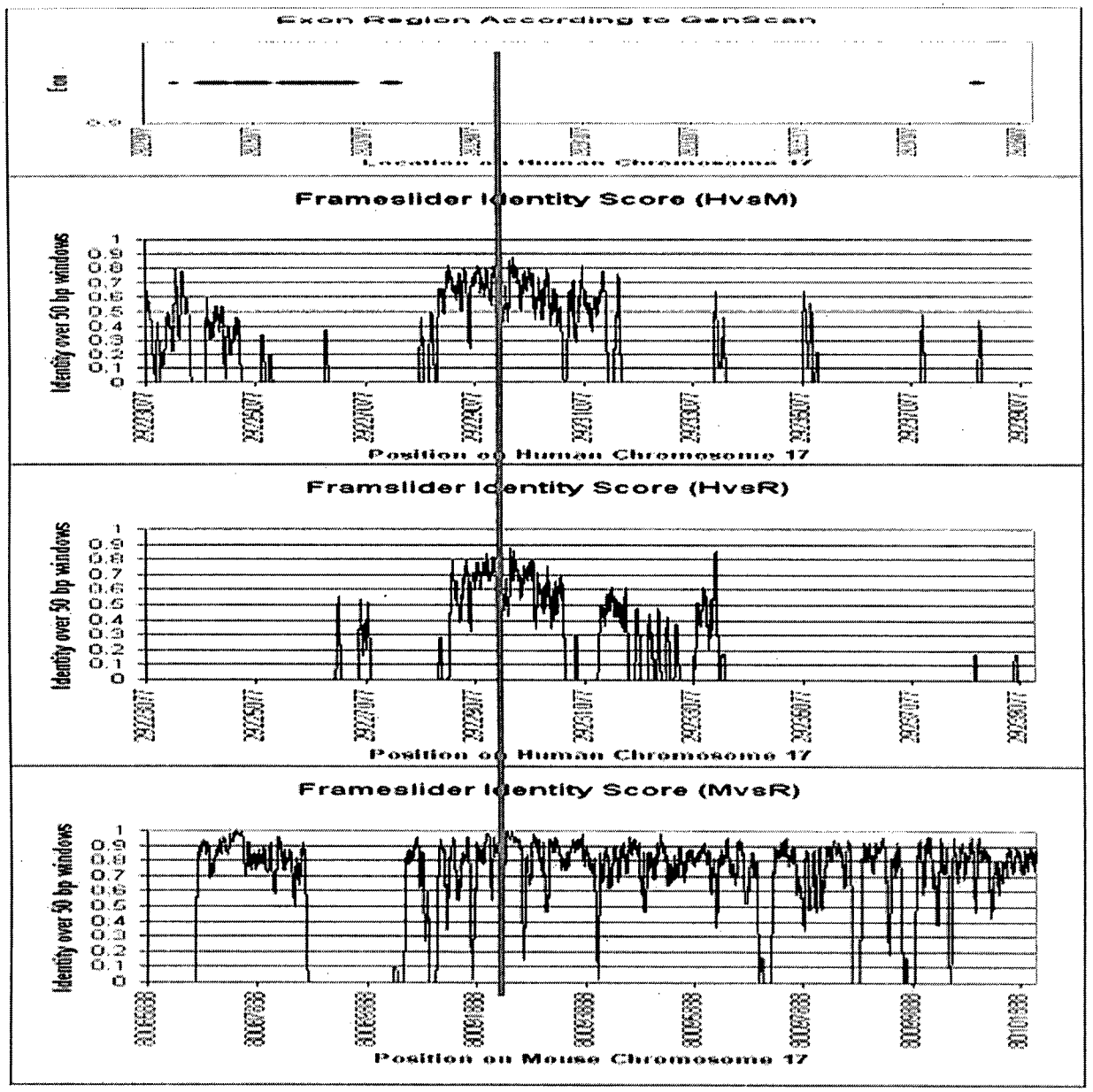


Figure 27. GenScan prediction for NT_010799.115 and the homology profile at Hchr17:29223077-29239353, Mchr11:80085888-80102169 and Rchr10:61719162-61735438. The top graph displays relative exon position, shown as blue dots. The vertical green line denotes the position of 5UR-HHR1.

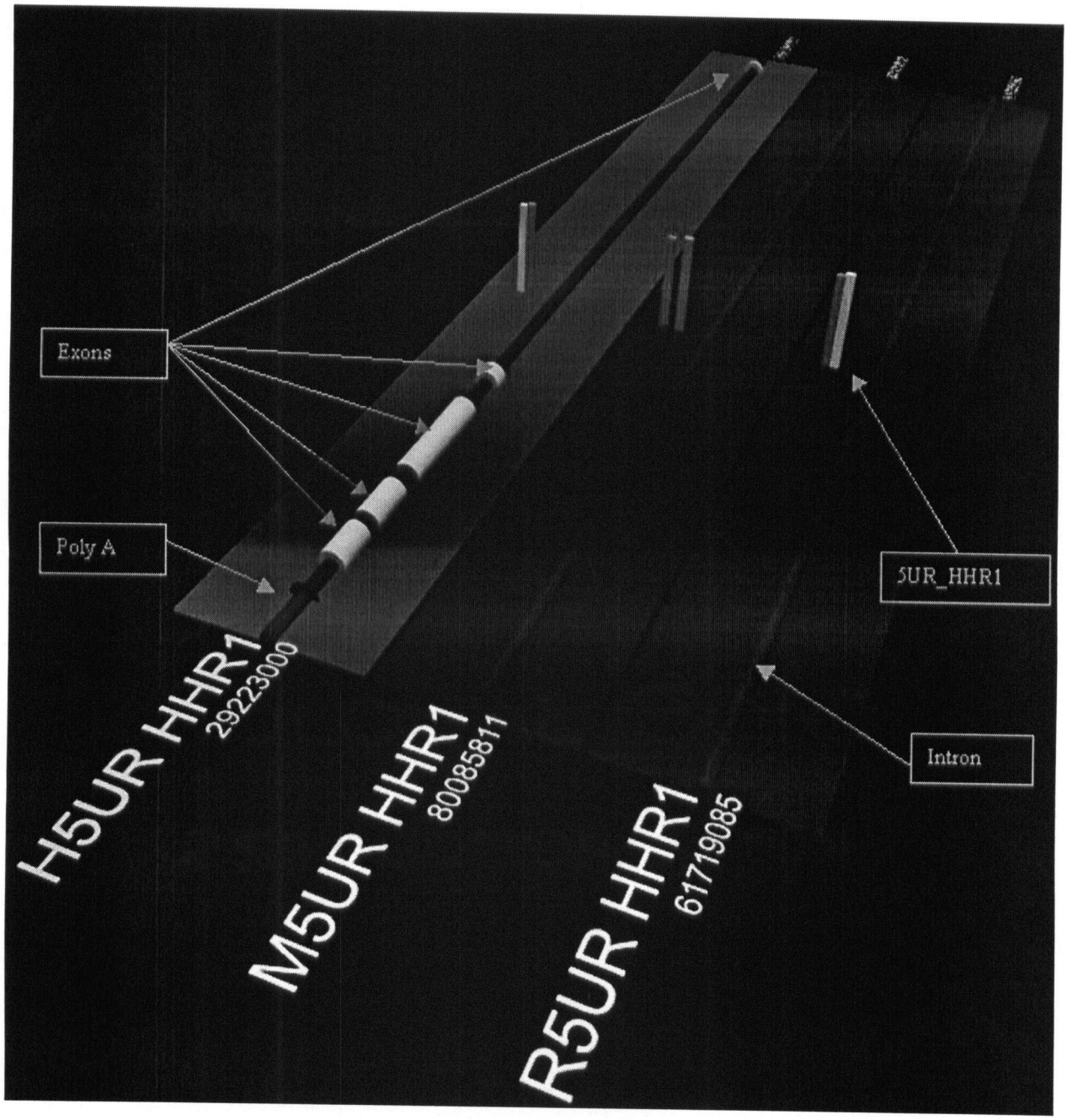


Figure 28. Sockeye presentation of 5UR-HHR1 and GenScan predictions for NT_010799.115. Blue rods denote non-coding sequences; purple rods denote exons on the reverse strand. The black semicircle denotes a poly A tail. Bars perpendicular to the plane are homologies at 5UR-HHR1 shown in the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. Note that the position of 5UR-HHR1 falls into an intron of the GenScan prediction.

Furthermore, 5UR-HHR1 is located in the intronic region, not in an exon of NT_010799.115. Therefore, 5HR_HHR1 is more likely to reflect a potential function other than protein coding.

3.7.2 5UR-HHR1 - Section 1b (H, M, & R VS F)

The *Fugu* 5UR was compared to 5UR-HHR1 of human, mouse, and rat using Pairwise BLAST. Three regions with exact matches common to human, mouse, rat and *Fugu* over at least 7 bp were found. The locations in *Fugu*, human, mouse, and rat are as follows: FCAAB01003481:21966-21973 (8 bp) with Hchr17:29229546-29229553, Mchr11:80092357-80092364, and Rchr10:61725631-61725638; FCAAB01003481:22591-22597 (7 bp) with Hchr17:29229549-29229555, Mchr11:80092360-80092366, and Rchr10:61725634-61725640; and FCAAB01003481:22060-22068 (9 bp) and Hchr17:29229590-29229598, Mchr11:80092363-80092371, and Rchr10:61725637-61725645. All of these matches have extremely high Expect value (>1393250), which means the matches are likely to be due to chance.

3.7.3 5UR-HHR1 – Section 2

The region extending from 100 bp upstream to 100 bp downstream of 5UR-HHR1 in humans and the corresponding regions in mouse and rat were downloaded. These are Hchr17:29229434-29229700, Mchr11:80092245-80092516 and Rchr10:61725519-61725786. These sequences and the whole *Fugu* 5UR were analyzed using MATCH™ and MatInspector with the settings listed in Chapter 2. Predictions that are shared between humans and at least one other species at the same aligned position are summarized in Table 7 and Table 8.

Table 7. Summary of MATCH™ predictions surrounding 5UR-HHR1 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, chr10 for rat, and contig CAAB01003481 for *Fugu*. Core S. = Core similarity. Matrix S. = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote tfactor binding sites that are detected with both minFN and minSUM settings. Bold characters denote the regions within 5UR-HHR1. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3; light blue, 4.

Tfactor	Human				Mouse				Rat				Pufferfish			
	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.
COMP1	29229475	29229498	0.786	0.661	80092264	80092287	0.786	0.619								
Elk-1	29229493	29229506	0.924	0.88	80092292	80092305	0.924	0.875	61725566	61725579	0.924	0.876				
COMP1	29229500	29229523	0.856	0.594	80092301	80092324	0.786	0.717								
CDP CR1	29229506	29229515	0.929	0.696	80092303	80092312	0.793	0.649								
Oct-1	29229510	29229524	0.824	0.571	80092309	80092323	0.824	0.717	61725583	61725597	0.824	0.713				
RFX1	**29229535**	**29229552**	**0.982**	**0.905**	**80092346**	**80092363**	**0.982**	**0.901**	**61725620**	**61725637**	**0.982**	**0.901**				
RFX1	**29229536**	**29229552**	**0.982**	**0.874**	**80092347**	**80092363**	**0.982**	**0.882**	**61725621**	**61725637**	**0.982**	**0.882**				
Oct-1	**29229539**	**29229553**	**0.888**	**0.711**	**80092350**	**80092364**	**0.888**	**0.715**	**61725624**	**61725638**	**0.888**	**0.715**				
c-Ets-1(p54)	**29229542**	**29229551**	**0.949**	**0.904**	**80092353**	**80092362**	**0.949**	**0.904**	**61725627**	**61725636**	**0.949**	**0.904**				
v-Myb	**29229545**	**29229554**	**0.904**	**0.847**	**80092356**	**80092365**	**0.904**	**0.847**					**724**	**733**	**0.904**	**0.837**
Pax-4	**29229558**	**29229578**	**0.817**	**0.619**	**80092370**	**80092390**	**0.794**	**0.688**								
HNF-4	**29229561**	**29229579**	**0.815**	**0.692**	***80092377***	**80092395**	***0.829***	***0.842***	***61725647***	**61725665**	***0.829***	***0.84***				
YY1	**29229568**	**29229587**	**1**	**0.862**	**80092384**	**80092403**	**1**	**0.853**	**61725654**	**61725673**	**1**	**0.853**				
COMP1	**29229569**	**29229592**	**0.914**	**0.615**	**80092385**	**80092408**	**0.914**	**0.625**	**61725655**	**61725678**	**0.914**	**0.625**				
COMP1	**29229581**	**29229604**	**0.786**	**0.568**	**80092397**	**80092420**	**0.786**	**0.613**	**61725667**	**61725690**	**0.786**	**0.585**				
COMP1	**29229585**	**29229608**	**0.786**	**0.568**	**80092401**	**80092424**	**0.786**	**0.561**	**61725671**	**61725694**	**0.786**	**0.586**				
CDP CR1	**29229587**	**29229596**	**0.781**	**0.759**	**80092403**	**80092412**	**0.781**	**0.759**	**61725673**	**61725682**	**0.781**	**0.759**				
AP-1	**29229588**	**29229598**	**0.829**	**0.848**	**80092404**	**80092414**	**0.829**	**0.848**	**61725674**	**61725684**	**0.829**	**0.848**	**818**	**828**	**0.955**	**0.965**
AP-1	**29229589**	**29229599**	**0.955**	**0.967**	**80092405**	**80092415**	**0.955**	**0.967**								
Pax-4	**29229589**	**29229609**	**0.764**	**0.59**	**80092405**	**80092425**	**0.764**	**0.589**	**61725675**	**61725695**	**0.764**	**0.635**				
Pax-4	29229617	29229637	0.771	0.643					61725703	61725723	0.795	0.599				
Pax-4	29229626	29229646	0.788	0.652	80092441	80092461	0.818	0.59	61725711	61725731	0.818	0.59				
Hand1/E47	29229638	29229653	0.868	0.859	80092450	80092465	1	0.908	61725720	61725735	1	0.908				
Pax-4	29229653	29229673	0.888	0.603	80092461	80092481	0.888	0.621	61725731	61725751	0.888	0.621				
HNF-4	29229671	29229689	0.883	0.632	80092479	80092497	0.755	0.606	61725749	61725767	0.848	0.681				
Oct-1	29229671	29229685	0.888	0.684	80092479	80092493	0.848	0.681	61725749	61725763	0.755	0.606				

Table 8. Summary of MatInspector predictions surrounding 5UR-HHR1 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, chr10 for rat, and contig CAAB01003481 for *Fugu*. Core S. = Core similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote tfactor binding sites that are identified when the core similarity setting was either 0.70 or 1.00. Bold characters denote the regions within 5UR-HHR1. Boxes highlighted in orange are tfactor predictions shared by 3 species; light blue, 4. All of the predictions shown in this table are within 5UR-HHR1.

Tfactor	Further Information	Human		Mouse		Rat		Fugu	
		Beginning	End	Beginning	End	Beginning	End	Beginning	End
V$IRFF/IRF3.01	*Interferon regulatory factor 3 (IRF-3)*	*29229544*	*29229558*	*80092355*	*80092369*	*61725629*	*61725643*	723	737
V$IRFF/IRF7.01	*Interferon regulatory factor 7 (IRF-7)*	*29229544*	*29229558*	*80092355*	*80092369*	*61725629*	*61725643*		
V$IRFF/IRF2.01	*Interferon regulatory factor 2*	*29229544*	*29229558*	*80092355*	*80092369*	*61725629*	*61725643*		
V$IRFF/IRF1.01	*Interferon regulatory factor 1*	*29229544*	*29229558*	*80092355*	*80092369*	*61725629*	*61725643*		
V$HAML/ AML3.01	*Runt-related transcription factor 2 / CBFA1 (core-binding factor, runt domain, alpha subunit 1)*	*29229575*	*29229589*	*80092391*	*80092405*	*61725661*	*61725675*		
V$PIT1/PIT1.01	*Pit1, GHF-1 pituitary specific pou domain transcription factor*	*29229588*	*29229598*	*80092404*	*80092414*	*61725674*	*61725684*		
V$AP1F/AP1.01	AP1 binding site	29229585	29229605	80092401	80092421	61725671	61725691	814	834

There is no tfactor binding site predicted by MATCH^{TM} on the forward strand for the human, mouse, or rat *NF1* 5UR under the minFP setting, although predictions do exist for the *Fugu* 5UR under this setting (Table 7). There are quite a few tfactor binding site predictions that are shared by human, mouse, and rat. Two predictions from regions in the *Fugu* sequence that are pinpointed by Pairwise BLAST are v-Myb, which is the oncogene of Avian myeloblastosis virus (AMV), and AP-1, which is Activator Protein. This AP-1 site is also worth noting because it is the only MATCH^{TM} prediction that is shared by all four species.

All MatInspector tfactor binding site predictions are either shared by human, mouse, and rat, or not shared at all (Table 8). Two of the predictions were also found in the *Fugu* 5UR sequence by Pairwise BLAST - Interferon Regulatory Factor 3 and AP-1 (Activator Protein). Since AP-1 is predicted by both MatInspector and MATCH^{TM} at the same position and is shared by all four species, this prediction is promising. These tfactor binding site predictions and the 5UR-HHR1 are shown by Sockeye in Figure 29.

3.7.4 5UR-HHR2 – Section 1a

5UR-HHR2 and 5UR-HHR3 are both located within 1488 bp upstream of the translation start site. Section 1a was done for 5UR-HHR2 and 5UR-HHR3 separately, using the human, mouse, and rat. Section 1b for 5UR-HHR2 and 5UR-HHR3 was replaced by a modified mVISTA and Frameslider analysis together with Pairwise BLAST using the 1488 bp segment upstream of the human, mouse, rat, and *Fugu* translation start sites. The analyses for transcriptional regulatory factors in Section 2 were then done for these 1488 bp segments that include both 5UR-HHR2 and 5UR-HHR3.

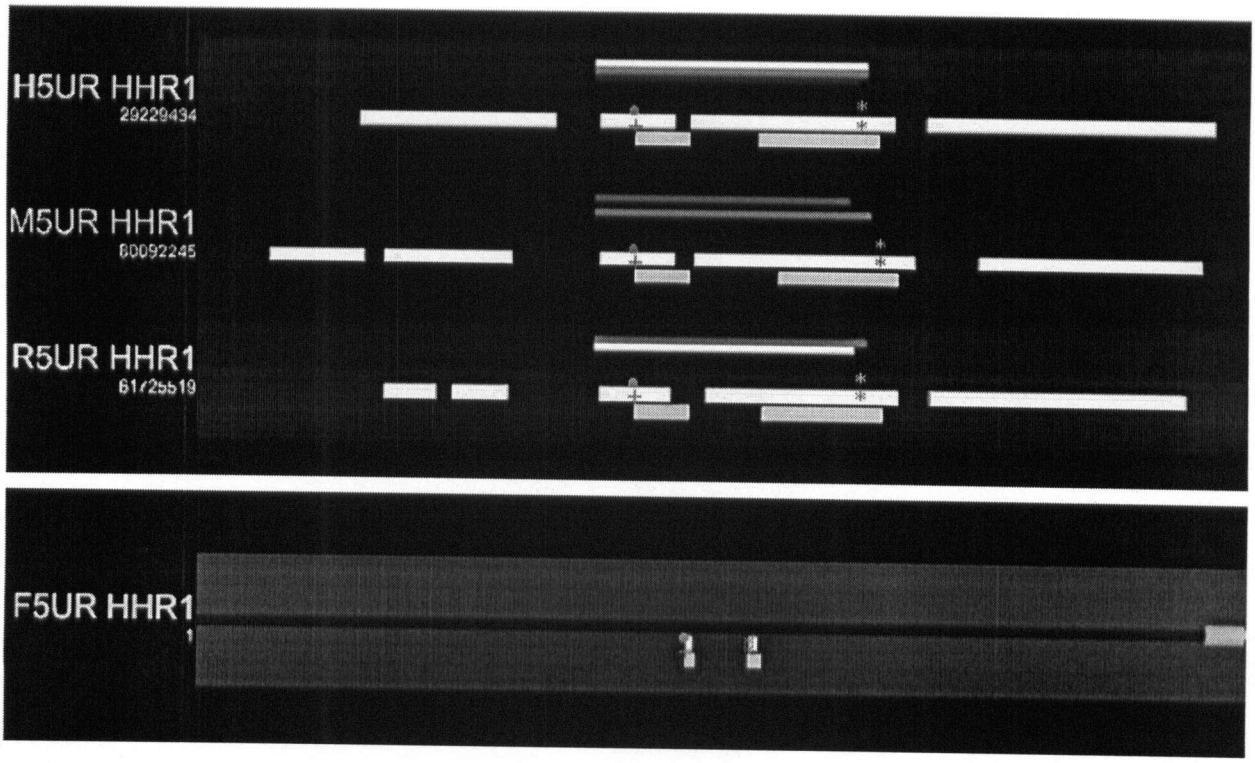
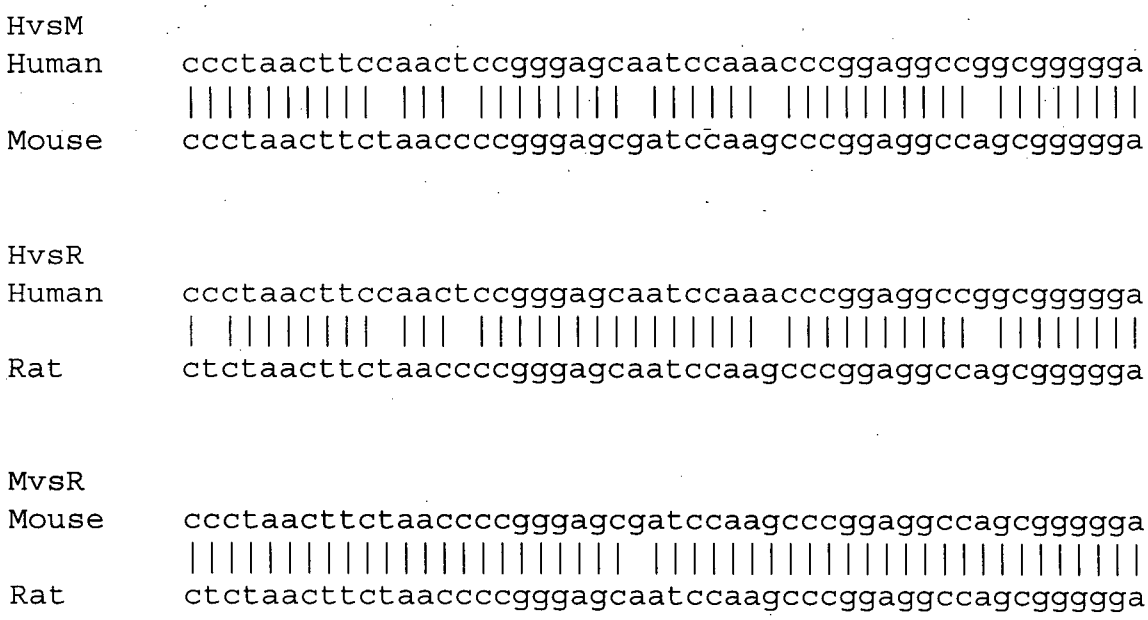


Figure 29. Sockeye presentation of tfactor predictions surrounding 5UR-HHR1 (Hchr17:29229434-29229700, Mchr11:80092245-80092516 and Rchr10:61725519-61725786) and *Fugu* 5UR (1488 bp). The mammalian species are shown in same scale, but the scale for *Fugu* 5UR is 5.59 times bigger. Blue rods denote non-coding regions; the purple rods denote exon 1 (in *Fugu*). The bars above non-coding regions represent the sequence of 5UR-HHR1 with homologies indicated by the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. Boxes below the non-coding regions are tfactor prediction from MATCH^{TM} (yellow) and MatInspector (orange). Only the position of tfactor predictions for v-Myb (•), IRF-3 (+), and AP-1 (*) are shown. The precise location of other tfactors can be found in Table 7 and Table 8.

a) mVista



b) Combined alignment

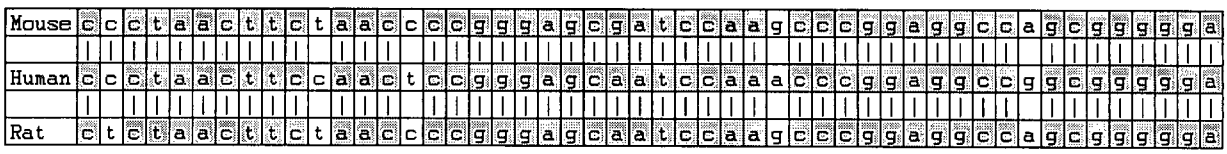


Figure 30. Alignment of 5UR-HHR2 in Hchr17:29271537-29271586, Mchr11:80124609-80124658, and Rchr10:61760457-61760506 from (a) mVista and (b) when they are combined. Nucleotides highlighted in yellow are shared by all three species.

5UR-HHR2 is a 50-bp window located at Hchr17:29271537-29271586, Mchr11:80124609-80124658, and Rchr10:61760457-61760506 (Figure 30). The identities are as follows: HvsM–0.90, HvsR–0.90, and MvsR–0.96. There is no repeat in this region, according to RepeatMasker. The positions of 5UR-HHR2 in human, mouse, and rat relative to their own translation start sites are -689 to –640, -683 to –634, and –511 to –462, respectively. Note that the major TSS for humans is at –484 relative to the translation start site.

According to UCSC, this region does not fall into any GenScan prediction in human or rat, but it does fall into GenScan prediction chr11_16.17 in mouse. Mchr11_16.17 is coded on the reverse strand, and it has supporting ESTs (Figure 31). The predicted mRNA of this predicted gene was BLASTed against the EST databases of human, mouse, rat, and *Fugu* using blastn with every parameter set as default except Expect, which was raised to 10 for lower stringency. Hits with low Expect values were found in mouse (8e-36 to 2e-18) and human (4e-36 to 4e-11), but not in rat, where the best hit had an Expected value of 1.0. No hit was found in *Fugu*.

It is unlikely that the homology is a result of the sequence lying in a coding sequence. All of the best hits with an Expect value smaller than 4e-11 in the human EST database were related to neurofibromin, due to an overlap of the predicted gene with ESTs of *NF1*. The best hits in the mouse EST database do not have annotation that is related to the ESTs of *NF1*. In addition, there were no convincing EST data from rat, and no mRNA was found in the *Fugu* EST database. Therefore, there is no evidence that gene Mchr11_16.17 exists in human, and this GenScan prediction is probably a false positive. For these reasons, no more analysis was done regarding this predicted gene in mouse. If this gene does exist in human, it is probably not located in the same region in relationship to NF1.

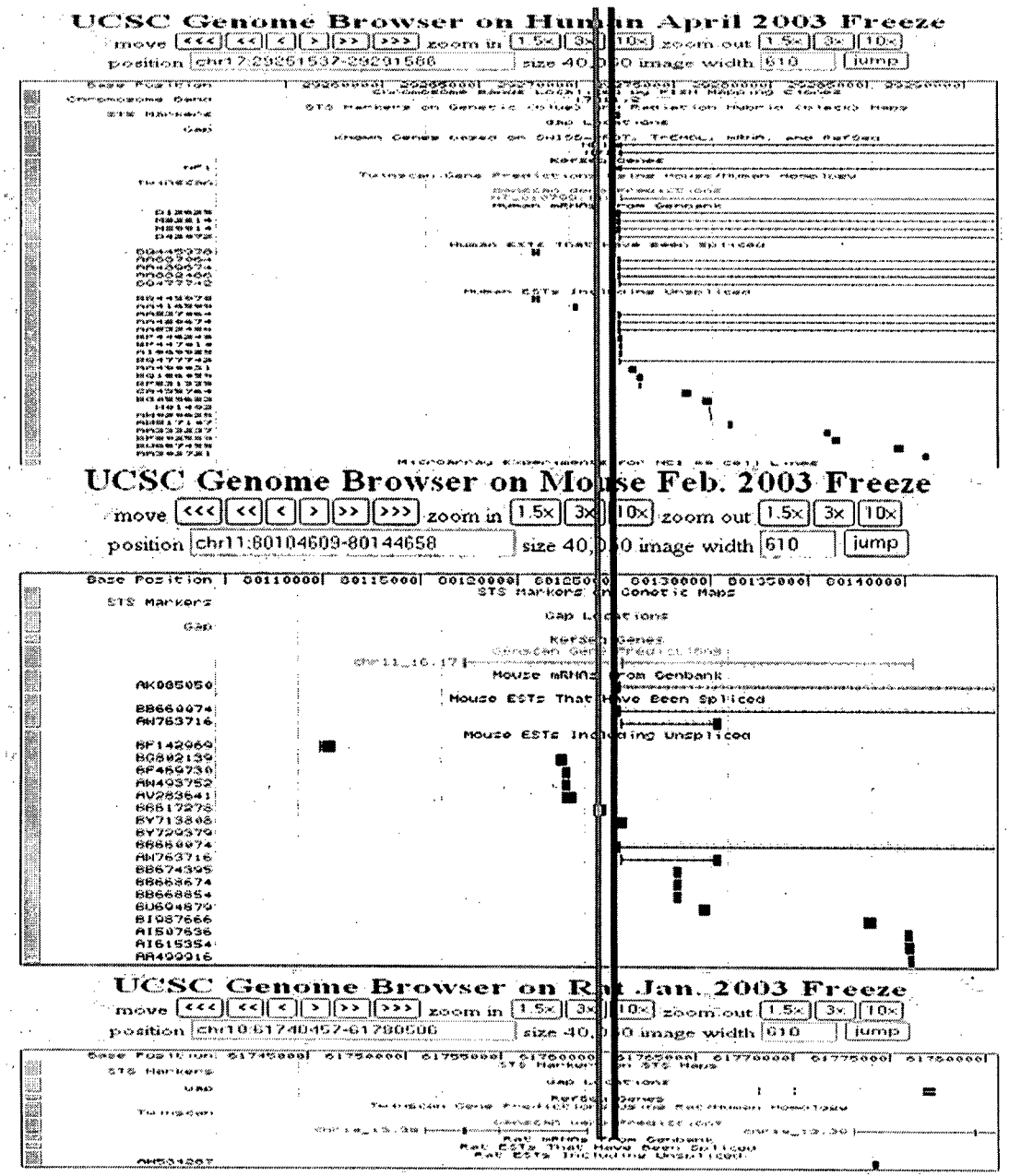


Figure 31. UCSC annotations at Hchr17:29251537-29291586, Mchr11:80104609-80144658, and Rchr10:61740457-61780506. The vertical green line denotes the location of 5UR-HHR2, the red line denotes the exact location of the TSS of the human *NF1* gene and the approximate locations of the TSSs of the mouse and rat *NF1*. Note the position of chr11_16.17, a GenScan predicted gene on the opposite strand in mouse.

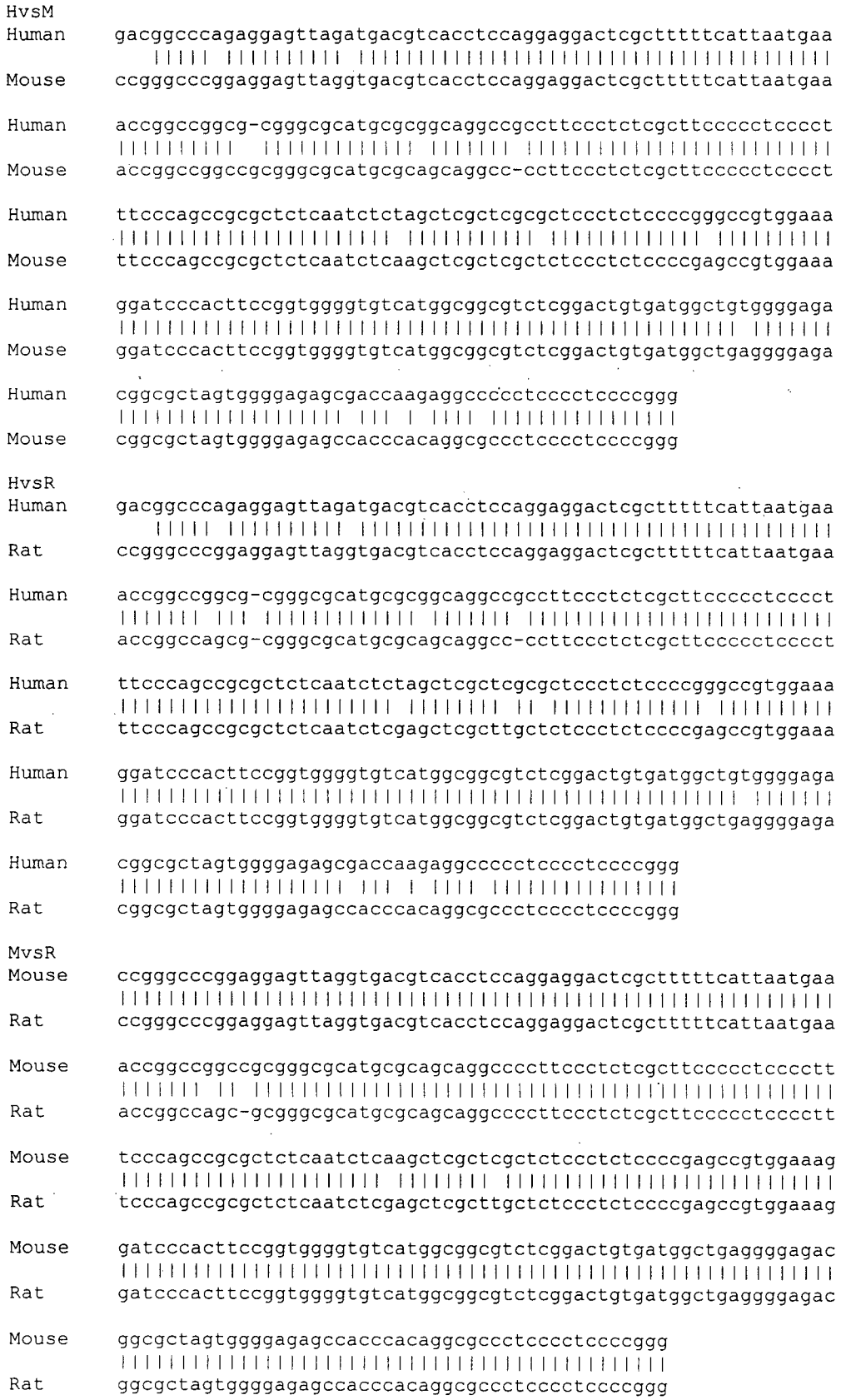
3.7.5 5UR-HHR3 – Section 1a

5UR-HHR3 is a long homologous region in human, mouse, and rat (Figure 32). The regions are located at Hchr17:29271707-29271993 (287 bp), Mchr11:80124780-80125066 (287 bp), and Rchr10:61760628-61760913 (286 bp). The ranges of identity of the windows are as followed: HvsM - 0.90 to 1, HvsR – 0.88 to 1, and MvsR – 0.96 to 1. The positions of 5UR-HHR3 of human, mouse, and rat relative to their own translation start sites are –519 to –233, -512 to –226, and –340 to –55. Therefore, this region spans the major TSS for human and mouse. Also, since there is high homology between mouse and rat, the difference in distances relative to the translation start site may indicate that that roughly 171 to 176 bp of genomic sequence are missing as a result of incomplete sequence around the NF1 translation start site and exon 1 in rat genome. RepeatMasker reported no repeat sequence in 5UR-HHR3 for the three species.

3.7.6 5UR-HHR2 & 5UR-HHR3 – Section 1b

The 1488 bp segments upstream of the translation start site for human, mouse, rat, and *Fugu* were downloaded as followed: Hchr17:29270738-29272225, Mchr11:80123804-80125291, Rchr10:61759480-61760967, and FCAAB01003481:21242-22729. The *Fugu* 5UR was used as the primary sequence and aligned against human, mouse, and rat using mVista. The alignment was then analyzed using Frameslider with a window size of 20 and cutoff values of 0.70 for FvsH and FvM, and 0.65 for FvsR. Note that although ORF homology for FvsR is 0.68, 0.65 was chosen as the cutoff because a window size of 20 leads to increments of 0.05 in identity. The following number of regions of high homology were found in these comparisons: FvsH – 8, FvsM – 7, FvsR – 16. Regions that were identical or overlapped in all three comparisons were identified – three such regions were found (Figure 33). As shown in the figure, some of these alignments are not tight, and there are extensive gaps in some cases. Furthermore, out of these three regions that show strong homology between the mammalian species and *Fugu*, only region

a) mVista


HvsM
Human gacggcccagaggagttagatgacgtcacctccaggaggactcgctttttcattaatgaa
Mouse ccgggcccggaggagttaggtgacgtcacctccaggaggactcgctttttcattaatgaa
Human accggccggcg-cgggcgcatgcgcggcaggccgccttccctctcgcttcccccctcccct
Mouse accggccggccgcgggcgcatgcgcagcaggcc-ccttccctctcgcttcccccctcccct
Human ttcccagccgcgctctcaatctctagctcgctcgcgctccctctccccgggccgtggaaa
Mouse ttcccagccgcgctctcaatctcaagctcgctcgctctccctctccccgagccgtggaaa
Human ggatcccacttccggtggggtgtcatggcggcgtctcggactgtgatggctgtggggaga
Mouse ggatcccacttccggtggggtgtcatggcggcgtctcggactgtgatggctgaggggaga
Human cggcgctagtggggagagcgaccaagaggcccctccccctccccggg
Mouse cggcgctagtggggagagccacccacaggcgccctccccctccccggg
HvsR
Human gacggcccagaggagttagatgacgtcacctccaggaggactcgctttttcattaatgaa
Rat ccgggcccggaggagttaggtgacgtcacctccaggaggactcgctttttcattaatgaa
Human accggccggcg-cgggcgcatgcgcggcaggccgccttccctctcgcttcccccctcccct
Rat accggccagcg-cgggcgcatgcgcagcaggcc-ccttccctctcgcttcccccctcccct
Human ttcccagccgcgctctcaatctctagctcgctcgcgctccctctccccgggccgtggaaa
Rat ttcccagccgcgctctcaatctcgagctcgcttgctctccctctccccgagccgtggaaa
Human ggatcccacttccggtggggtgtcatggcggcgtctcggactgtgatggctgtggggaga
Rat ggatcccacttccggtggggtgtcatggcggcgtctcggactgtgatggctgaggggaga
Human cggcgctagtggggagagcgaccaagaggcccctccccctccccggg
Rat cggcgctagtggggagagccacccacaggcgccctccccctccccggg
MvsR
Mouse ccgggcccggaggagttaggtgacgtcacctccaggaggactcgctttttcattaatgaa
Rat ccgggcccggaggagttaggtgacgtcacctccaggaggactcgctttttcattaatgaa
Mouse accggccggccgcgggcgcatgcgcagcaggcccctttccctctcgcttcccccctcccctt
Rat accggccagc-gcgggcgcatgcgcagcaggcccctttccctctcgcttcccccctcccctt
Mouse tcccagccgcgctctcaatctcaagctcgctcgctctccctctccccgagccgtggaaag
Rat tcccagccgcgctctcaatctcgagctcgcttgctctccctctccccgagccgtggaaag
Mouse gatcccacttccggtggggtgtcatggcggcgtctcggactgtgatggctgaggggagac
Rat gatcccacttccggtggggtgtcatggcggcgtctcggactgtgatggctgaggggagac
Mouse ggcgctagtggggagagccacccacaggcgccctccccctccccggg
Rat ggcgctagtggggagagccacccacaggcgccctccccctccccggg

b) Combined Alignment

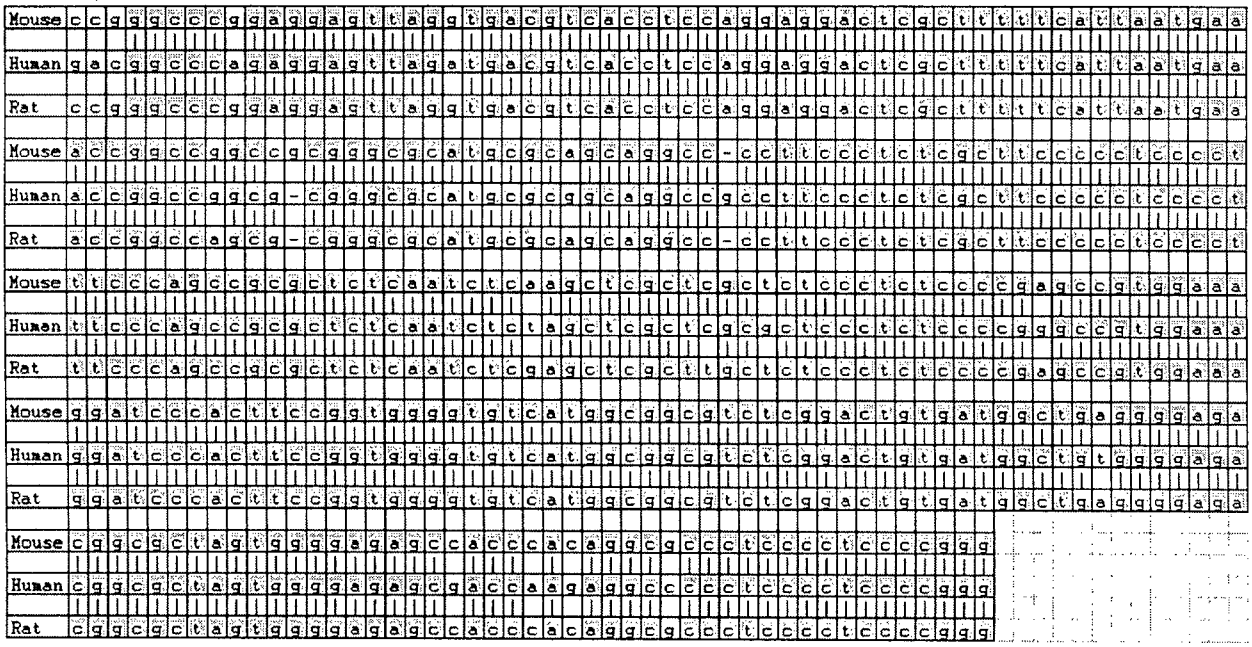


Figure 32. Alignment of 5UR-HHR3 in Hchr17:29271707-29271993 (287 bp), Mchr11:80124780-80125066 (287 bp), and Rchr10:61760628-61760913 from (a) mVista and (b) when they are combined. Nucleotides highlighted in yellow are shared by all three species.

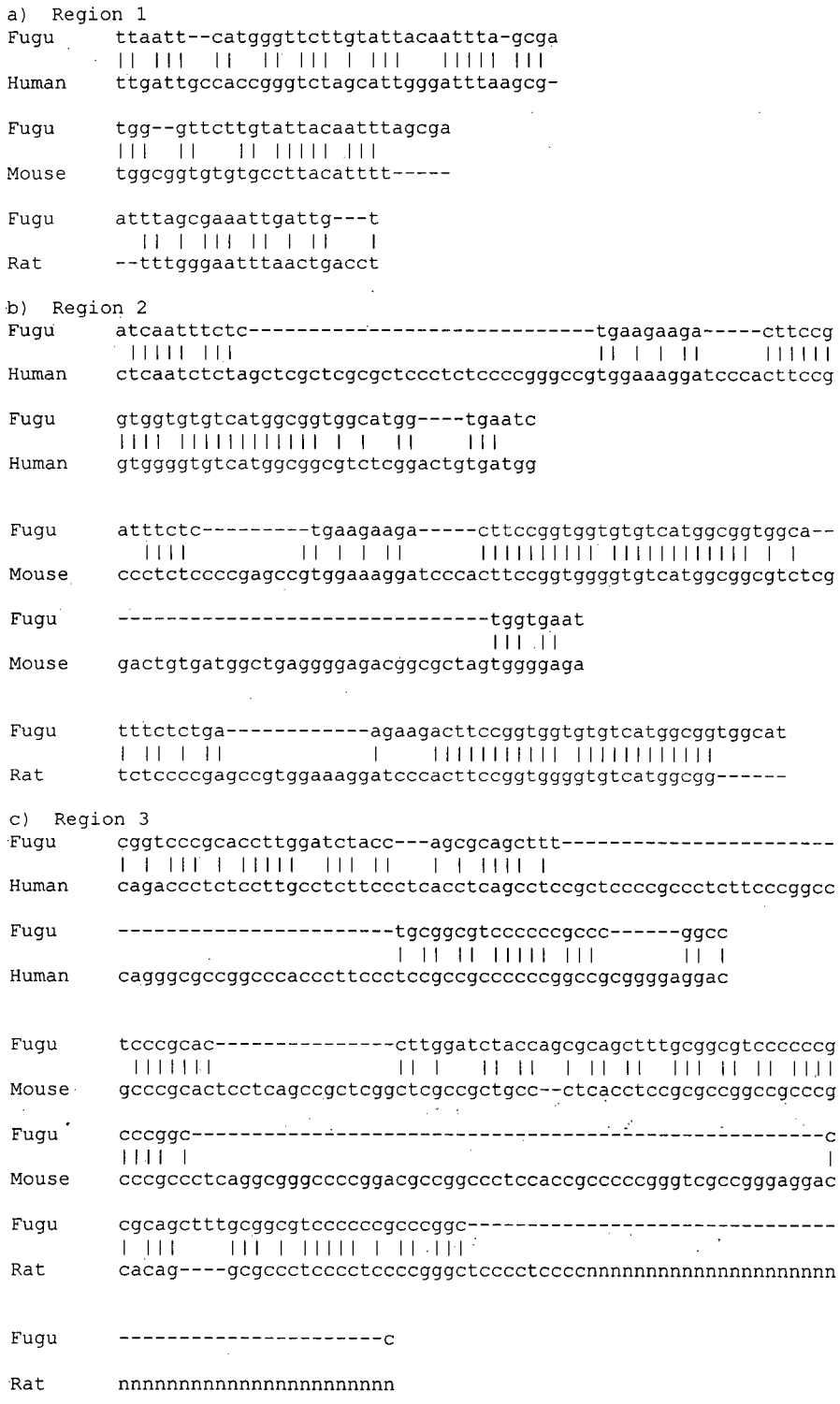


Figure 33. Alignments of the 1488 bp segments upstream of the *NF1* ORF in human, mouse, rat, and *Fugu*. Regions that meet the cutoff values from Frameslider and are shared by all four species (a, b, and c) are shown. Note that FvsR in region 1 only overlaps with FvsH and FvsM in the beginning. Also, in region 2, there are 24 bp (shown in blue) that are nearly identical in all 4 species. The extensive gaps that exist in region 3 indicate an unpromising alignment.

2 lies within a mammalian highly homologous region (5UR-HHR3). Interestingly, there is 24-bp-segment that is identical among human, mouse, and rat and varies by only 1 single bp in pufferfish. This sequence is **acttccggtggggtgtcatggcgg**. Frameslider indicates the identity over this range to be 0.85-0.95. This sequence, which is analyzed in more detail below, will be referred to as the NF1 5' Highly Conserved Sequence (NF1HCS). It is located at Hchr17:29271893-29271916, Mchr11:80124966-80124989, Rchr10:61760813-61760836, and F CAAB01003481:22551-22574.

Next, the *Fugu NF1* 5UR was compared to 5UR-HHR2 and 5UR-HHR3 of human, mouse, and rat using Pairwise BLAST. No exact match over least 7 bp or more was found that was common to human, mouse, rat and *Fugu* in 5UR-HHR2. 16 exact matches of at least 7 bp were found that were shared by to humans, mice, rats, and *Fugu* within 5UR-HHR3. A portion of NF1HCS is the only match that could be detected when the Wordsize was increased to 12 bp. When the same experiment was repeated while gradually lowering the 'Expect' value, NF1HCS was detected with an Expect value of 5.3. The other segments needed a much higher Expect value (>3000000). This means that the homology for NF1HCS is the least likely to be due to chance among the hits that were observed.

3.7.7 5UR-HHR2 – Section 2

Tfactor detection was performed in the region surrounding 5UR-HHR2 and 5UR-HHR3, with special attention to the region covered by NF5HCS. Regions extending from 100 bp upstream to 100 bp downstream of 5UR-HHR2 in human, mouse, and rat (Hchr17:29271437-29271686, Mchr11:80124509-80124758, and Rchr10:61760357-61760606) were analyzed using MATCH^{TM} and MatInspector (Table 9, Table 10). Only predictions that were shared between

Table 9. Summary of MATCH™ predictions surrounding 5UR-HHR2 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, and chr10 for rat. Core S. = Core similarity. Matrix S. = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote tfactor binding sites that are detected with both minFN and minSUM settings. Bold characters denote the regions within 5UR-HRR2. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Human				Mouse				Rat			
	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.
Evi-1	29271453	29271467	0.763	0.626	80124513	80124527	0.842	0.714	61760360	61760374	0.842	0.714
Pax-4	29271455	29271475	0.796	0.668	80124515	80124535	0.888	0.626	61760362	61760382	0.888	0.626
c-Ets-1(p54)	29271482	29271491	0.974	0.9					61760399	61760408	1	0.918
HNF-4	29271490	29271508	0.752	0.619	80124562	80124580	0.752	0.658	61760409	61760427	0.752	0.661
HNF-4	*29271510*	29271528	*0.883*	0.833	80124582	80124600	0.883	0.659	61760430	61760448	0.883	0.658
c-Ets-1(p54)	***29271568***	29271587	***1***	***0.949***	***80124640***	80124659	***1***	***0.949***	***61760488***	61760507	***1***	***0.949***
Hand1/E47	29271632	29271647	1	0.913	80124711	80124726	1	0.878	61760559	61760574	1	0.878
CP2	29271653	29271663	1	0.838	80124729	80124739	0.974	0.826	61760577	61760587	0.974	0.829

Table 10. Summary of MatInspector predictions surrounding 5UR-HHR2 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, and chr10 for rat. Core S. = Core similarity. Matrix S. = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote tfactor binding sites identified when the core similarity setting was either 0.70 or 1.00. Bold characters denote the regions within 5UR-HHR2. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Further Information	Human				Mouse				Rat			
		Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.
V$PBXF/PBX1.01	*Homeo domain factor Pbx-1*	*29271555*	*29271567*	*1*	*0.786*					61760475	61760487	1	0.784
V$AP2F/AP2.01	*Activator protein 2*	*29271566*	*29271578*	*1*	*0.9*	*80124638*	*80124650*	*1*	*0.917*	61760486	61760498	1	0.917

human and at least one other species at the same aligned position are summarized in the tables. Note that tfactor prediction for this region is done for *Fugu* in Section 2 of 5UR-HHR1 and some of the results are reported in Table 7 and 8.

MATCH^{TM} predicted no tfactor binding site on the forward strand for human, mouse, or rat under the minFP setting. Several tfactor binding site predictions were found that were common to human, mouse, and rat using the minFN setting. However, only c-ETS-1(p54), which is involved in mesodermal cell development during organ formation and tissue modeling, lies within 5UR-HHR2. This prediction appears with the minSUM setting for all three mammalian species.

All MatInspector tfactor binding site predictions shared by human and at least one other species lie within 5UR-HHR2, and all have a core similarity of 1.00. The predictions are quite different from those of MATCH^{TM}. However, the MATCH^{TM} c-ETS-1(p54) prediction and MatInspector AP-2 prediction are at the same position, which may mean that a tfactor binding site does exist but the identity of that tfactor is uncertain. The tfactor binding site predictions related to 5UR-HHR2 are illustrated by Sockeye in Figure 34.

3.7.8 5UR-HHR3 – Section 2

Regions extending from 200 bp (instead of 100 bp) upstream to 200 bp downstream of 5UR-HHR3 were downloaded as followed: Hchr17:29271507-29272193, Mchr11:80124580-80125266, and Rchr10:61760428-61760967. This longer region was chosen to provide an opportunity to compare the predictions from MATCH^{TM} and MatInspector to the ones that have previously been reported (Hajra *et al.*, 1994). Note that in rat the translation start site is

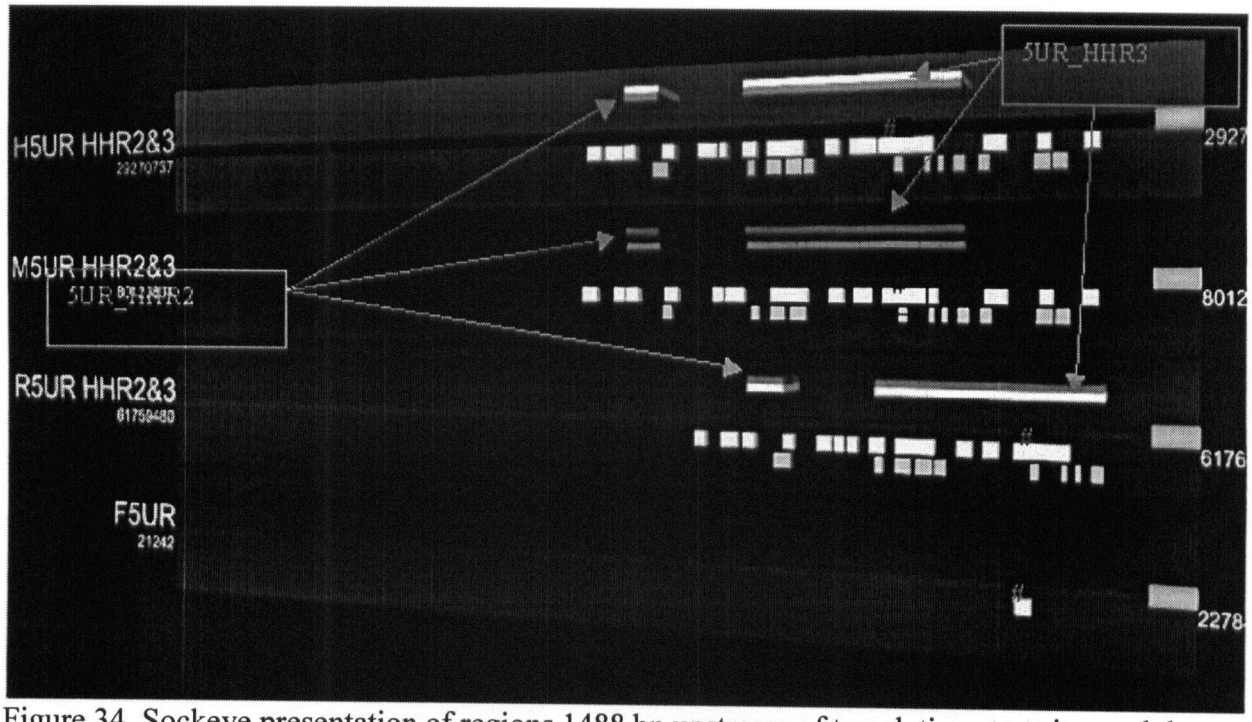


Figure 34. Sockeye presentation of regions 1488 bp upstream of translation start sites and the first 60 translated bp for human, mouse, rat, and *Fugu*. Blue rods denote non-coding regions; purple rods denote exons; black rods denote NF1 Highly Conserved Sequence. Bars above non-coding regions represent 5UR-HHR2 and 5UR-HHR3 with homologies indicated by the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. Boxes below are tfactor predictions from MATCH™ (yellow) and MatInspector (pink). Only the positions of tfactor predictions for Pax-4 (#) from MATCH™ are shown. The precise location of other tfactor predictions can be found in Tables 9, 10, 11, and 12.

annotated only 55 bp downstream of 5UR-HHR3, probably as a result of the incomplete sequence around the translation start site of the *NF1* gene. These sequences were analyzed using MATCH^{TM} and MatInspector (Table 11 and Table 12). Only predictions that were shared between the human sequence and that of at least one other species at the same aligned positions are summarized in the tables. Note that some of the tfactor predictions for this region in *Fugu* were reported above in Table 7 and 8.

MATCH^{TM} predicted two or three tfactor binding sites on forward strand for human, mouse, and rat under minFP settings. Many tfactor binding site predictions were shared by human, mouse, and rat at the minFN settings (Table 11). Of the 9 predictions shared between humans and mice that have previously been published (Hajra *et al.*, 1994), only CREB at Hchr17:29271726 was seen in this analysis, probably due to the use of a different program and an updated TRANSFAC library. Two tfactor predictions were found in the NF1HCS region - Pax-6 and Pax-4. The prediction for Pax-4 (Paired box-4), which is involved in cell fate, early patterning, and organogenesis, is especially important because it is common to human, mouse, rat, and *Fugu*.

MatInspector gave several predictions that are shared by humans and other species (Table 12). Most of these conserved predictions are within 5UR-HHR3. Although the predictions are quite different from those of MATCH^{TM}, both programs predict the previously-described CREB site in the same position. MatInspector also predicts the previously-described AP-2 site (Hajra *et al.*, 1994). The positions of the two previously-described Sp-1 sites (Hajra *et al.*, 1994) are predicted to be Hchr17:29271580-29271594 and Hchr17:29271599-29271613, but these two predictions were not found for mouse or rat. Although an Erythroid krueppel-like factor (EKLF) binding site shared by all three mammalian species was predicted within the NF1HCS, this prediction was different from the Pax-6 and Pax-4 predictions from MATCH^{TM}. Lastly, MatInspector

Table 11. Summary of MATCH^{TM} predictions surrounding 5UR-HHR3 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, chr10 for rat, and contig CAAB01003481 for *Fugu*. Core S. = Core similarity. Matrix S. = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote tfactor binding sites that are detected with both the minFN and minFP settings. Bold characters denote region within 5UR-HHR3. White characters denote previously reported tfactor binding sites. Blue characters denote the tfactor predicitons within NF1HCS. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3; light blue, 4.

Tfactor	Human				Mouse				Rat				Pufferfish			
	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.
HNF-4	29271510	29271528	0.883	0.833	80124582	80124600	0.883	0.659	61760430	61760448	0.883	0.656				
c-Ets-1(p54)	29271568	29271577	1	0.949	80124640	80124649	1	0.949	61760488	61760497	1	0.949				
Pax-6	29271622	29271642	0.831	0.688					61760535	61760555	0.925	0.608				
Pax-4	29271629	29271649	0.789	0.64					61760535	61760555	0.879	0.638				
Hand1/E47	29271632	29271647	1	0.913	80124711	80124726	1	0.878								
CP2	29271653	29271662	1	0.838	80124729	80124738	0.974	0.826	61760577	61760586	0.974	0.829				
Pax-4	29271686	29271706	0.881	0.737	80124736	80124756	0.789	0.668	61760607	61760627	0.881	0.66				
v-Maf	**29271721**	**29271739**	**0.91**	**0.696**	**80124794**	**80124812**	**0.87**	**0.673**	**61760642**	**61760660**	**0.87**	**0.673**				
HLF	**29271721**	**29271730**	**0.872**	**0.81**	**80124794**	**80124803**	**0.872**	**0.834**	**61760642**	**61760651**	**0.872**	**0.834**				
CREB	**29271723**	**29271734**	**1**	**0.989**	**80124796**	**80124807**	**1**	**0.992**	**61760644**	**61760655**	**1**	**0.992**				
ATF	**29271724**	**29271737**	**1**	**0.973**	**80124797**	**80124810**	**1**	**0.98**	**61760645**	**61760658**	**1**	**0.98**				
AP-1	**29271725**	**29271735**	**0.935**	**0.864**	**80124798**	**80124808**	**0.935**	**0.883**	**61760646**	**61760656**	**0.935**	**0.883**				
CREB	**29271725**	**29271736**	**1**	**0.926**	***80124798***	**80124809**	***1***	***0.955***	**61760646**	**61760657**	**1**	**0.955**				
CREB	**29271725**	**29271736**	**1**	**0.97**	**80124798**	**80124809**	**1**	**0.994**	**61760646**	**61760657**	**1**	**0.994**				
CRE-BP1	**29271725**	**29271736**	**1**	**0.961**	***80124798***	**80124809**	***1***	***0.987***	***61760646***	**61760657**	**1**	***0.987***				
CRE-BP1/c-Jun	**29271727**	**29271734**	***1***	***1***	***80124800***	**80124807**	***1***	***1***	***61760648***	**61760655**	***1***	***1***				
CREB	**29271727**	**29271734**	**1**	**1**	***80124800***	**80124807**	***1***	***1***	***61760648***	**61760655**	***1***	***1***				
Pax-4	**29271727**	**29271747**	**0.788**	**0.622**	**80124800**	**80124820**	**0.788**	**0.622**	**61760648**	**61760668**	**0.788**	**0.622**				
c-Ets-1(p54)	**29271738**	**29271747**	**0.974**	**0.901**	**80124811**	**80124820**	**0.974**	**0.901**	**61760659**	**61760668**	**0.974**	**0.901**				
HNF-3beta	**29271747**	**29271761**	**0.93**	**0.855**	**80124820**	**80124834**	**0.93**	**0.855**	**61760668**	**61760682**	**0.93**	**0.855**				
Pax-4	**29271751**	**29271771**	**0.817**	**0.602**	**80124824**	**80124844**	**0.817**	**0.602**	**61760672**	**61760692**	**0.817**	**0.602**				
CDP,CR1	**29271757**	**29271766**	**0.794**	**0.753**	**80124830**	**80124839**	**0.794**	**0.753**	**61760678**	**61760687**	**0.794**	**0.753**				
v-Myb	**29271763**	**29271772**	**0.91**	**0.835**	**80124836**	**80124845**	**0.91**	**0.835**	**61760684**	**61760693**	**0.91**	**0.835**				
Pax-6	**29271802**	**29271822**	**0.812**	**0.616**	**80124875**	**80124895**	**0.812**	**0.616**	**61760722**	**61760742**	**0.812**	**0.616**				
Pax-4	**29271836**	**29271856**	**0.803**	**0.643**	**80124909**	**80124929**	**0.803**	**0.649**	**61760756**	**61760776**	**0.803**	**0.673**				
Pax-6	**29271852**	**29271872**	**0.812**	**0.557**	**80124915**	**80124935**	**0.832**	**0.578**								
HNF-4	**29271875**	**29271893**	**0.883**	**0.765**	**80124948**	**80124966**	**0.883**	**0.768**	**61760795**	**61760813**	**0.883**	**0.768**				
Oct-1	**29271875**	**29271889**	**0.888**	**0.785**	**80124948**	**80124962**	**0.888**	**0.785**	**61760795**	**61760809**	**0.888**	**0.785**				
CDP,CR1	**29271883**	**29271892**	**0.768**	**0.782**	**80124956**	**80124965**	**0.768**	**0.782**	**61760803**	**61760812**	**0.768**	**0.782**				
CDP,CR3+HD	**29271883**	**29271892**	**0.8**	**0.838**	**80124956**	**80124965**	**0.8**	**0.838**	**61760803**	**61760812**	**0.8**	**0.838**				
Pax-6	**29271890**	**29271910**	**0.791**	**0.555**	**80124963**	**80124983**	**0.791**	**0.555**	**61760810**	**61760830**	**0.791**	**0.555**				
Pax-4	**29271903**	**29271923**	**1**	**0.758**	**80124976**	**80124996**	**1**	**0.758**	**61760823**	**61760843**	**1**	**0.758**	**22561**	**22581**	**1**	**0.704**
Oct-1	**29271922**	**29271935**	**0.964**	**0.886**	**80124995**	**80125008**	**0.964**	**0.886**	**61760842**	**61760855**	**0.964**	**0.886**				
Pax-4	**29271924**	**29271944**	**0.902**	**0.691**	**80124997**	**80125017**	**0.902**	**0.691**	**61760844**	**61760864**	**0.902**	**0.691**				
COMP1	**29271924**	**29271947**	**0.786**	**0.593**	**80124997**	**80125020**	**0.786**	**0.616**	**61760844**	**61760867**	**0.786**	**0.616**				
Pax-6	**29271924**	**29271944**	**0.767**	**0.568**	**80124997**	**80125017**	**0.767**	**0.666**	**61760844**	**61760864**	**0.767**	**0.666**				
Staf	29272012	29272032	0.926	0.842	80125081	80125101	0.926	0.819								
Oct-1	29272022	29272036	0.883	0.729	80125091	80125105	0.883	0.584								
HNF-4	29272023	29272041	0.811	0.626	80125092	80125110	0.811	0.62								
RFX1	29272080	29272097	0.982	0.872	80125151	80125168	0.982	0.871								
N-Myc	29272082	29272093	0.889	0.904	80125153	80125164	0.889	0.904								
Oct-1	29272084	29272098	0.776	0.661	80125155	80125169	0.776	0.66								
USF	29272084	29272091	0.871	0.897	80125155	80125162	0.871	0.897								
Ik-1	29272086	29272098	1	0.96	80125157	80125169	1	0.957								
Pax-4	29272139	29272159	0.789	0.584	80125205	80125225	0.788	0.623								

Table 12. Summary of MatInspector predictions surrounding 5UR-HHR3 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, and chr10 for rat. C = Core similarity. M = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote predictions that are identified when the core similarity setting is 1.00. Bold characters denote the regions within 5UR-HHR3. Pink characters denote previously reported tfactor binding site predictions. Blue characters denote predictions in NF1HCS. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Further Information	Human				Mouse				Rat			
		Beginning	End	C	M	Beginning	End	C	M	Beginning	End	C	M
V$PBXF/ PBX1.01	*Homeo domain factor Pbx-1*	*29271565*	*29271567*	*1*	*0.786*					*61760475*	*61760487*	*1*	*0.784*
V$AP2F/ AP2.01	*Activator protein 2*	*29271566*	*29271578*	*1*	*0.9*					*61760486*	*61760498*	*1*	*0.917*
V$WHZF/ WHN.01	*Winged helix protein involved in hair keratinization and thymus epithelium differentiation*	*29271694*	*29271704*	*1*	*0.974*	*80124767*	*80124777*	*1*	*0.952*	*61760615*	*61760625*	*1*	*0.952*
V$CREB/ CREBP1CJUN.01	***CRE-binding protein 1/c-Jun heterodimer***	***29271721***	***29271741***	***1***	***1***	***80124794***	***80124814***	***1***	***1***	***61760642***	***61760662***	***1***	***1***
V$CREB/ CREBP1.01	***cAMP-responsive element binding protein 1***	***29271721***	***29271741***	***1***	***0.861***	***80124794***	***80124814***	***1***	0.861	***61760642***	***61760662***	***1***	***0.861***
V$CREB/ CREB.01	***cAMP-responsive element binding protein***	***29271721***	***29271741***	***1***	***1***	***80124794***	***80124814***	***1***	***1***	***61760642***	***61760662***	***1***	***1***
V$CREB/ ATF6.02	***Activating transcription factor 6, member of b-zip family, induced by ER stress***	***29271721***	***29271741***	***1***	***0.872***	***80124794***	***80124814***	***1***	0.882	***61760642***	***61760662***	***1***	***0.882***
V$CREB/ CREB.02	***cAMP-responsive element binding protein***	***29271721***	***29271741***	***1***	***0.992***	***80124794***	***80124814***	***1***	0.997	***61760642***	***61760662***	***1***	***0.997***
V$CREB/ CREB.03	***cAMP-response element-binding protein***	***29271721***	***29271741***	***1***	***0.915***	***80124794***	***80124814***	***1***	0.949	***61760642***	***61760662***	***1***	***0.949***
V$CREB/ ATF.02	***ATF binding site***	***29271721***	***29271741***	***1***	***0.964***	***80124794***	***80124814***	***1***	0.968	***61760642***	***61760662***	***1***	***0.968***
V$CREB/ CREB.04	***cAMP-response element binding protein***	***29271721***	***29271741***	***1***	***0.962***	***80124794***	***80124814***	***1***	0.993	***61760642***	***61760662***	***1***	***0.993***
V$CREB/ CREBP1.02	***CRE-binding protein 1***	***29271721***	***29271741***	***1***	***0.951***	***80124794***	***80124814***	***1***	0.984	***61760642***	***61760662***	***1***	***0.984***
V$CREB/ ATF.01	***activating transcription factor***	***29271721***	***29271741***	***1***	***0.956***	***80124794***	***80124814***	***1***	0.968	***61760642***	***61760662***	***1***	***0.968***
V$CREB/ VJUN.01	***v-Jun***	***29271721***	***29271741***	***1***	***0.921***	***80124794***	***80124814***	***1***	0.873	***61760642***	***61760662***	***1***	***0.873***
V$E4FF/ E4F.01	***GLI-Krueppel-related transcription factor, regulator of adenovirus E4 promoter***	***29271726***	***29271738***	***1***	***0.977***	***80124799***	***80124811***	***1***	0.983	***61760647***	***61760659***	***1***	***0.983***
V$HOXF/ HOX1-3.01	***Hox-1.3, vertebrate homeobox protein***	***29271750***	***29271766***	***1***	***0.836***	***80124823***	***80124839***	***1***	0.836	***61760671***	***61760687***	***1***	***0.836***
V$PAX2/ PAX2.01	***Zebrafish PAX2 paired domain protein***	***29271748***	***29271770***	***1***	***0.816***	***80124821***	***80124843***	***1***	0.816	***61760669***	***61760691***	***1***	***0.816***
V$EGRF/ WT1.01	**Wilms Tumor Suppressor**	**29271773**	**29271787**	**1**	**0.905**					**61760694**	**61760708**	**1**	**0.905**
V$EGRF/ EGR2.01	**Egr-2/Krox-20 early growth response gene product**	**29271773**	**29271787**	**1**	**0.792**					**61760694**	**61760708**	**1**	**0.79**
V$EGRF/ EGR1.01	***Egr-1/Krox-24/NGFI-A immediate-early gene product***	**29271773**	***29271787***	**1**	***0.805***					***61760694***	***61760708***	**1**	***0.805***
V$EGRF/ EGR3.01	**early growth response gene 3 product**	**29271773**	**29271787**	**1**	**0.818**					***61760694***	***61760708***	**1**	**0.821**
V$EKLF/ EKLF.01	***Erythroid krueppel like factor (EKLF)***	***29271897***	***29271907***	***1***	***0.945***	***80124970***	***80124980***	***1***	***0.945***	***61760817***	***61760827***	***1***	***0.945***
V$MZF1/ MZF1.01	***MZF1***	***29271937***	***29271943***	***1***	***1***	***80125010***	***80125016***	***1***	***0.985***	***61760857***	***61760863***	***1***	***0.985***
V$MZF1/ MZF1.01	***MZF1***	***29271954***	***29271960***	***1***	***1***	***80125027***	***80125033***	***1***	***1***	***61760874***	***61760880***	***1***	***1***
V$ZBPF/ ZF9.01	**Core promoter-binding protein (CPBP) with 3 Krueppel-type zinc fingers**	**29271974**	**29271988**	**1**	**0.888**	**80125047**	**80125061**	**1**	**0.888**	**61760894**	**61760908**	**1**	**0.888**
V$ZBPF/ ZBP89.01	*Zinc finger transcription factor ZBP-89*	*29272006*	*29272020*	*1*	*1*	*80125076*	*80125090*	*1*	*0.978*				
V$ZBPF/ ZBP89.01	*Zinc finger transcription factor ZBP-89*	*29272076*	*29272090*	*1*	*0.966*	*80125147*	*80125161*	*1*	*0.966*				
V$ZBPF/ ZF9.01	Core promoter-binding protein (CPBP) with 3 Krueppel-type zinc fingers	29272076	29272090	1	0.888	80125147	80125161	1	0.921				
V$HIFF/ HIF1.02	*Hypoxia inducible factor, bHLH / PAS protein family*	*29272080*	*29272092*	*1*	*0.945*	*80125151*	*80125163*	*1*	*0.945*				
V$IKRS/ IK1.01	*Ikaros 1, potential regulator of lymphocyte differentiation*	*29272086*	*29272098*	*1*	*0.943*	*80125157*	*80125169*	*1*	*0.937*				
V$MINI/ MUSCLE_INI.01	Muscle Initiator Sequence	29272101	29272119	1	0.865	80125172	80125190	1	0.865				

failed to detect any tfactor binding site in *Fugu* NF1HCS that was conserved with the human, mouse, and rat. The positions of NF1HCS, 5UR-HHR2, 5UR-HHR3 and the tfactor predictions surrounding these highly homologous regions are shown by Sockeye in Figure 34.

3.8 Summary for 5UR

Three highly homologous regions were found in NF1 5UR (locations within brackets are positions relative to the translation start site in each species). 5UR-HHR1 was found at Hchr17:29229534-29229600 (-42692 to –42626), Mchr11:80092345-80092416 (-32947 to –32876) and Rchr10:61725619-61725686 (-35349 to -35282). 5UR-HHR2 was found at Hchr17:29271537-29271586 (-689 to –640), Mchr11:80124609-80124658 (-683 to –634), and Rchr10:61760457-61760506 (-511 to –462). 5UR-HHR3 was found at Hchr17:29271707-29271993 (-519 to –233), Mchr11:80124780-80125066 (-512 to -226), and Rchr10:61760628-61760913 (-340 to -55).

Within 5UR-HHR1, the most important tfactor prediction was an AP-1 site predicted at Hchr17:29229585-29229605 (-42641 to -42621), Mchr11:80092401-80092421 (-32891 to -32871), Rchr10:61725671-61725691 (-35297 to –35277), and FCAAB01003481:22055-22075 (-675 to –655) by MATCH^{TM} and MatInspector. This site is shared by human, mouse, and rat at the same aligned position and possibly also by *Fugu* if the Pairwise BLAST alignment is correct.

Within 5UR-HHR2, the most noteworthy spot is at Hchr17:29271566-29271578 (-660 to -648), Mchr11:80124638-80124650 (-654 to –642), and Rchr10:61760486-61760498 (-482 to –470). c-

Ets-1 (p54) is predicted at this location is by MATCH^{TM}, but AP-2 is predicted by Matinspector at the same location.

Within 5UR-HHR3, the most important discovery is NF1HCS, a 24 bp segment that is completely identical among humans, mice, and rats, and that differs by only 1 bp in *Fugu*. This sequence is located at Hchr17:29271893-29271916 (-333 to –310), Mchr11:80124966-80124989 (-326 to –303), Rchr10:61760813-61760836 (-155 to –132), and F CAAB01003481:22551-22574 (-179 to –156). MATCH^{TM} predicts a binding site for Pax-4, a tfactor involved in cell fate, early patterning, and organogenesis, in the NF1HCS of all four species. No prediction from MatInspector is shared by all four species in the NF1HCS region. Other important results include confirmation of the previously-reported CREB site by MATCH^{TM} and MatInspector, and of the previously-described AP-1 and Sp1 sites in the human by MatInspector.

3.9 Analyses for NF1HCS

Although NF1HCS was previously recognized as part of the homologous segment of the 5UR in human and mouse (Hajra *et al.*, 1994), it has never been defined separately or associated with a potential function. Because NF1HCS is so highly conserved among human, mouse, rat, and pufferfish, additional analyses were done to explore the possibility that NF1HCS might function as the *NF1* core promoter element. A search was performed for other instances of this 24 bp mammalian sequence within the genomes of different organisms, with special attention to the relative position and relationship between NF1HCS and adjacent genes. The sequence as also tested against the Eukaryotic Promoter Database and TFDD to check whether NF1HCS had been reported in association with the promoter regions of other genes.

3.9.1 The Occurrence of NF1HCS in Various Genomes

If NF1HCS contains a core promoter element, it would be expected to occur in association with other genes and to be widely conserved by evolution. However, an exact match of all 24 bp would not be expected in other locations in the genome because core promoter elements are usually less than 24 bp long. The mammalian NF1HCS sequence acttccggtgggggtgtcatggcgg was BLASTed against the complete NCBI database using default settings except that Expect was increased to 1000000 and all filters were disabled. Hits were found in human, mouse, rat, pufferfish, *Drosophila* and some bacteria (*e.g.* Thermosynechococcus elongatus, Mycobacterium tuberculosis). Hits were also obtained in the separated genomic searches of the *Caenorhabditis elegans* and *Saccharomyces cerevisiae* genomic databases.

If NF1HCS were a core promoter sequence for other genes, it would be expected to exhibit three characteristics: First, it should occur on the same strand as the gene. Second, it should not lie in the coding sequence of the gene. Third, it should be upstream of the ORF. To see whether NF1HCS fullfiled these expectations, genomic BLAST was done on the human, mouse, rat, pufferfish, and fruitfly genomic sequences individually. The mammalian NF1HCS was used as the query and default settings were used except Expect value was raised to 10. BLAT from UCSC was used to pinpoint the locations of hits found in the mammalian species. If NF1HCS was found within an annotated gene, the NF1HCS sequence was checked against the mRNA of that gene using Pairwise BLAST to determine the relationship to the coding region. BLAT is not available for *Fugu* or *Drosophila*, and no annotation is available for the *Fugu* genome. Therefore, GenScan was used to predict any potential genes and their distance relative to hits in *Fugu*. NCBI annotations for the *Drosophila* genome were used.

1000 bp upstream of the ORF of Drosophila NF1 homologue was downloaded from ENSEMBL (chr3R:21797380-21798379) and compared to NF1HCS using Pariwise BLAST. No exact match of 7 bp (minimum wordsize setting) was found.

The results are summarized in Table 13. There are three interesting findings. First, portions of NF1HCS were found in other organisms in various locations. The portion of NF1HCS that was found to be homologous varied somewhat but usually included the nucleotides gtgtcatggcgg near the 3' end of the sequence. Second, most hits occurred in the vicinity of an annotated gene or a gene prediction, but most of these genes or predictions were on the opposite strand from the gene except in *Fugu*. Third, the hits were usually within a gene rather than upstream of it. Two hits were in an exon of a gene on the same strand in *Fugu*, but the other hits in mammalian species and Drosophila were found within an intron. Overall, these results do not support to the possiblity that NF1HCS contains a core promoter element.

3.9.2 Comparison with Eukaryotic Promoter Database (EPD)

2994 promoter regions (499 bp upstream and 100 bp downstream of the TSS) of genes with known translation start sites from various organisms were downloaded. Within this dataset, 1871 regions were human, 196 mouse, 119 rat, 120 *Drosophila*, but none were from *Fugu*. The total number of nucleotides for all the regions was 1796400 bp.

If NF1HCS includes a core promoter element, NF1HCS or a portion of it would be expected to occur within the promoter region of other genes and to lie downstream of the transcription start site, as it does in *NF1*. In fact, a location around 181 bp downstream of the TSS would be

Table 13. Summary of human, mouse, rat, pufferfish, and fruitfly genomic BLASTs with mammalian NF1HCS (acttccggtggggtgtcatggcgg) as the query sequence. Chr = chromosome, except for pufferfish, where the contig is given. For Str, + means the alignment is on the same strand as the annotated or predicted gene, - means the alignment is on the opposite strand. 'Begin' and 'End' represent the corresponding positions on the chromosome. Expect is the expect values indicating a stronger alignment. Letters in red means that the alignment is found within an exon of a gene located on the same strand.

Organism	Chr	Str	Begin	End	Expect	Comment
Human	17	+	29271893	29271916	7.00E-05	Whole NF1HCS is found upstream of the *NF1* gene
	8	-	140817722	140817738	1	Last 17 bp of NF1HCS is found within intron of the gene *MGC4737* on the other strand
	9	-	110308656	110308671	4	Last 16 bp of NF1HCS is found within intron of the gene *Loc113220* on the other strand
	9	+	92524373	92524388	4	Last 16 bp of NF1HCS is found within the intron of the Twinscan prediction chr9.93.004.a on the other stand
	2	-	202013354	202013369	4	Last 16 bp of NF1HCS is found within the intron of the gene *CASP10* on the other stand
	1	+	221235854	221235869	4	Last 16 bp of NF1HCS is found within the intron of the gene *FLJ38993* on the other stand
Mouse	11	+	80124966	80124989	6.00E-05	Whole NF1HCS is found upstream of the *NF1* gene
	7	+	21694794	21694810	0.94	Region between the 6th bp and the 22nd bp of NF1HCS is found within the intron of the 2700043M03R gene on the other strand
	17	-	45955809	45955824	3.7	Region between the 7th bp and the 22nd bp of NF1HCS is found within the exon of the Snt2-pending geneon the other strand
	14	+	79055923	79055938	3.7	Region between the 7th bp and the 22nd bp of NF1HCS is found within the intron of the GenScan prediction chr14_16.8 on the other strand
	1	-	124337184	124337199	3.7	Region between the 7th bp and the 22nd bp of NF1HCS is found within the intron of GenScan prediction chr1_25.6 on the other strand
Rat	10	+	61760813	61760836	7.00E-05	Whole NF1HCS is found upstream of the *NF1* gene
	16	+	66545237	66545252	3.9	Region between the 6th bp and the 21st bp of NF1HCS is found within the intron of the TwinScan prediction chr16_957.1 on the other strand
	9	-	13111180	13111195	3.9	Region between the 7th bp and the 22nd bp of NF1HCS does not have any annotated gene or predictions within 1000 bp upstream and 1000 bp downstream
	5	+	116411580	116411595	3.9	Region between the 7th bp and the 22nd bp of NF1HCS is found within the intron of the GenScan prediction chr5_24.17 on the same strand
	4	+	73851825	73851840	3.9	Region between the 7th bp and the 22nd bp of NF1HCS does not have any annotated gene or predictions within 1000 bp upstream and 1000 bp downstream
	2	-	29135008	29135023	3.9	Region between the 5th bp and the 20th bp of NF1HCS does not have any annotated gene or predictions within 1000 bp upstream and 1000 bp downstream
Pufferfish	CAAB0100 3481.1	+	22551	22574	0.002	Whole NF1HCS except the 12th bp is found upstream of the *NF1* gene
	CAAB0100 0706.1	-	69159	69173	2.1	Region between the 9th bp and the 23rd bp of NF1HCS is found within the exon of a GenScan prediction on the same strand.
	CAAB0100 0628.1	+	31587	31601	2.1	Region between the 4th bp and the 18th bp of NF1HCS is found within the intron of a GenScan prediction on the same strand.
	CAAB0100 0553.1	+	54484	54498	2.1	Region between the 6th bp and the 20th bp of NF1HCS is found 1033 bp upstream of the promoter of a GenScan prediction on the same strand.
	CAAB0100 3944.1	-	11321	11334	8.3	Region between the 8th bp and the 21st bp of NF1HCS is found within the exon of a GenScan prediction on the same strand.
	CAAB0100 2786.1	+	23059	23072	8.3	Region between the 3th bp and the 16th bp of NF1HCS is found within the intron of a GenScan prediction on the same strand.

	CAAB0100 1499.1	+	59515	59528	8.3	Region between the 3th bp and the 16th bp of NF1HCS is not found within or upstream of any GenScan prediction
	CAAB0100 1086.1	+	5476	5489	8.3	Region between the 10th bp and the 23nd bp of NF1HCS is found within the intron of a GenScan prediction on the same strand.
	CAAB0100 0944.1	+	26949	26962	8.3	Region between the 9th bp and the 22nd bp of NF1HCS is found within the intron of a GenScan prediction on the same strand.
	CAAB0100 0418.1	-	29670	29657	8.3	Region between the 6th bp and the 19th bp of NF1HCS is found within the intron of a GenScan prediction on the same strand.
	CAAB0100 0176.1	-	65763	65750	8.3	Region between the 4th bp and the 17th bp of NF1HCS is found within the intron of a GenScan prediction on the other strand.
	CAAB0100 0084.1	-	52620	52607	8.3	Region between the 4th bp and the 17th bp of NF1HCS is found within the exon of a GenScan prediction on the other strand.
Fruitfly	3R	+	17440679	17440698	0.21	Region between the 3rd bp and the 22nd bp of NF1HCS is found within the exon of gene *E2f*on the other strand.
	X	+	17101536	17101550	0.84	Region between the 7th bp and the 21st bp of NF1HCS does not have any annotated gene or predictions within 1000 bp upstream and 1000 bp downstream
	X	-	14912942	14912955	3.3	Region between the 9th bp and the 22nd bp of NF1HCS does not have any annotated gene or predictions within 1000 bp upstream and 1000 bp downstream
	3R	+	18877437	18877450	3.3	Region between the 4th bp and the 17th bp of NF1HCS is found 400 bp upstream of annotated gene *BcDNA:LD21504* on the other strand.
	3L	+	4375093	4375106	3.3	Region between the 10th bp and the 23rd bp of NF1HCS is found within the annotated gene *CG7447*on the same strand.

expected, because most core promoters (*e.g.* TATA box, DPE) have a consistent distance and direction from the TSS (Kadonaga, 2002). Unfortunately, EPD only includes the first 100 bp downstream of the TSS, so searching this database for NF1HCS does not provide information on whether this sequence occurs in the expected relationship to the TSS of these genes.

The promoter regions within EPD were compared to NF1HCS with Pairwise BLAST. With the default settings (with filters off and the forward strand chosen), no perfect match for all 24 bp was found. When the settings were adjusted to look for longest exact match, 2239 perfect matches of 7 bp or more were found. The distribution was as follows: 1624 matches of 7 bp, 412 matches of 8 bp, 102 matches of 9 bp, 30 matches of 10 bp, and 7 matches of 11 bp. The longest matches were 13 bp, and there were only two of them. The first of these included the first 13 bp of NF1HCS and was located 24 bp upstream of the TSS of the human gene Ovarian cancer overexpressed 1 (NCBI reference NM_015945) on chromosome 20. The second match included the last 13 bp of NF1HCS and was located 42 bp downstream of the TSS of the human gene Ubiquitin specific protease 5 (NCBI reference NM_003481) on chromosome 12. 1413 of the 2994 promoter regions in EPD had more then one match of >7bp with NF1HCS.

There were three interesting results from this analysis. First, a portion of NF1HCS was found in 47% of the promoter regions. Second, the two longest matches of 13 bp were both from human promoter regions. Third, portions of NF1HCS were found both upstream and downstream of the TSS. This experiment showed that portion of NF1HCS could be found in a large portion of promoter regions, although its location relative to the TSS was not fixed.

3.9.3 Comparison with the TRRD Database

There are many subsets of the TRRD database. TRRDUNITS4, which contains information on transcription factor binding sites, eukaryotic gene promoters, enhancers, transcription regulatory regions, gene expression regulation, and corresponding bibliography references, was used for this study. Default settings were used but Filter was turned off and Gap alignment was enabled. This database has a total of 971 sequences.

When the mammalian NF1HCS was used as the query sequence, a segment from the 7th through the 17th bp was aligned to the reverse strand of a promoter region (TRRD accession number: P00656) upstream of the TSS of the mitochondrial glycerol-3-phosphate acyltransferase (GPAT) gene in mouse.

When the *Fugu* NF1HCS was used as the query sequence, five alignments were found.

1. The 4th through the 15th bp were aligned to the reverse stand of a promoter region (TRRDR accession number P0013) upstream of the TSS of lipoxygenase 1 (LoxA) in barley (*Hordeum vulgare L*).
2. The 12th through the 22nd bp were aligned to the forward strand of an enhancer region (TRRD accession P00598) 3000 bp upstream of the ORF of the granulocyte/macrophage colony stimulating factor (GM-CSF) gene in human.
3. The 11th through the 21st bp were aligned to the forward strand of a silencer region (TRRD accession number P00842) upstream of the TSS of the lactoferrin (LFER) gene in human.
4. The 8th through the 18th bp were aligned to the forward strand of a promoter region (TRRD accession number P00543) upstream of the TSS of the acyl-coenzyme A synthetase gene (ACS) in rat.

5. The 11th through the 21st bp were algined to the forward strand of a promoter region upstream (TRRD accession number P00470) of the TSS of the cellular retinol-binding protein II (CRBPII) gene in mouse.

Because NF1HCS is on the same strand as NF1, alignments that involve the reverse strand are probably not relevant to the regulation of *NF1* transcription. The TRRD alignments involving the forward strand indicate that a portion of NF1CS1 occurs in genomic sequences involved in transcriptional regulation of human, mouse, and rat.

3.9.4 Potential RNA Structure

The whole mammalian NF1HCS was searched for potential RNA structure with Rfam and yielded no result. A 64 bp segment that includes NF1HCS and extends 20 bp upstream and downstream was then used as a query in Rfam. Again, no result was obtained. Thus, there is no indication that NF1HCS has a recognized secondary RNA structure.

Because of the relatively simple but rigid search function in SCOR, the entire mammalian NF1HCS sequence was first input, and then various portions of this sequence were searched. Searches were done of all substrings made by removing one base at a time from the 3' end of NF1HCS, then of all substrings made by removing one base at a time from the 5' end of NF1HCS, and finally of all substrings made by serially removing one base from each end of NF1HCS at the same time. No hits were obtained with NF1HCS or with any of these substrings that were more than 4 bp long. Thus, SCOR did not identify any 3D structure for the RNA produced by NF1HCS.

3.9.5 Comparison NF1HCS with promoter regions of other genes

NF1HCS is 24 bp long, but most core promoter sequences are only 6 bp long. Therefore, if NF1HCS includes a core promoter element, the segment of complete sequence identity among the mammalian species and of >95% identity with *Fugu* almost certainly includes surrounding sequence in addition to the core promoter element itself. Therefore, I examined the regions surrounding known core promoter elements of other genes to see whether similar high levels of identity are observed in the homologous human, mouse, rat, and *Fugu* genes.

Five genes with defined core promoter elements were chosen for analysis. They are the beta-globin (*HBB*), alpha-skeletal actin 1 (*ACTA1*), transcription factor AP-2 gamma (*TFAP2C*), TATA box-binding protein-associated factor (*TAF7*), and lymphocyte-specific protein-tyrosine kinase (*LCK*) genes. *HBB* and *ACTA1* were chosen because they contain a putative TATA box in *Fugu* (Gillemans *et al.*, 2002; Vekatesh *et al.*, 1996). *TFAP2C* was chosen because it has been found to have an Inr element in human (Hasleton *et al.*, 2003), and TAF7 was selected because it has been shown to have Inr and DPE elements in human (Zhou *et al.* 2001). Inr and DPE have not been identified in *Fugu* genes. *LCK* was included because this gene has been found to have two promoters in human and mouse, and both are known to share a highly conserved 11 bp segment with the homologous gene in *Fugu* (Brenner *et al.*, 2002). *LCK* has no known TATA or Inr.

Nucleotides upstream and downstream of each core promoter element were downloaded so that the whole region was roughly 50 bp long. This was followed by mVISTA manual alignment and frameslider identity calculation using a 24 bp window, which is the length of NF1HCS.

The TATA box and the surrounding region of the *HBB* gene are summarized in Table 14. The TATA box associated with this gene in human (Lewis *et al.*, 2000), mouse (Jacob *et al.*, 1994), and *Fugu* (Gillemans *et al.*, 2002) do not have the usual TATAAA consensus sequence. The homologous rat TATA box was found by manual inspection. As shown in Table 14, the region upstream of the TATA box exhibits only high homology between human, mouse and rat, and much less homology to *Fugu.* According to UCSC, the *HBB* gene has no upstream CpG island in human, mouse, and rat.

The TATA box and surrounding region of *ACTA1* gene are summarized in Table 15. The TATA box in human, mouse, rat were found by manual inspection. They all have the TATAAA consensus sequence and lie 30 bp upstream of the major transcription initiation site in human, 31 bp in mouse, and about 39 bp in rat. The *Fugu* TATA box also has the usual consensus sequence (Venkatesh *et al.*, 1996). As shown in Table 15, the region surrounding the TATA box exhibits high homology among human, mouse, and rat. The homology between human and *Fugu* is also quite high, considering the cDNA of the *NF1* gene is also only 0.70 identity between mammals and *Fugu* (Table 5). According to UCSC, there is a 3477 bp CpG island that begins 843 bp upstream of the TSS in human, but there is no upstream CpG island in mouse and rat.

The Inr box and surrounding region of the *TFAP2C* gene are summarized in Table 16. The Inr core element that has been described in human (Hasleton *et al.* 2003) was found in the homologous mouse and rat genes by manual inspection. This Inr was situated at the right place with respect to major TSS in human (Hasleton *et al.*, 2003). However, it was 4 bp further upstream than expected in mouse according UCSC. Because the TSS of rat has not been defined, the location of Inr with respect to the TSS was not clear. There has been no published research

Table 14. Comparison of regions surrounding the TATA box of the beta-globin (*HBB*) gene in human, mouse, rat, and *Fugu*. Str = Strand. The *Fugu* sequence was obtained from Gillemans *et al.*, 2002. The alignment between human and mouse, rat, and pufferfish is shown. Nucleotides identical to human are highlighted in yellow. Bold capital letters are nucleotides for the TATA box. The region between the two red lines in each alignment is the 24 bp with the highest identity. Identity, shown in the last column, is calculated by dividing the number of nucleotides identical to human in each species by the length of the region (24 bp).

Species	TATA box			Alignment Surrounding TATA box	Identity
	Str	Location	Sequence		
Human	-	chr11:5207364-5207369	ATAAAA	NA	NA
Mouse	-	chr7:92296946-92296951	ATATAA	agggcaggagccagggctgggcATAAAAgtcagggcagagccatctattg agggcaggagccagggcagagcATATAAggtgaggtaggatcagttgctc	0.92
Rat	-	chr1:152536246-152536250	ATAAAA	-aggcaggagccagggctgggcATAAAAgtcagggcagagccatctattg gagagcaggagccaggcagagcATAAAAggtgggcggagatcagtcgctc	0.88
Fugu	-	Scaffold_628:102043-102049	AATTAAA	agg-gcaggagccagggctgggcATAAA---Agtcagggcagagccatctattg ggtgggggtgtctcggcacAATTAAAtagacttcagctc----ccactgcaac	0.54

Table 15. Comparison of regions surrounding the TATA box of the alpha-skeletal actin 1 (*ACTA1*) gene in human, mouse, rat, and *Fugu*. Str = Strand. The *Fugu* sequence was obtained from Venkatesh *et al.*, 1996. The alignment between human and mouse, rat, and pufferfish is shown. Nucleotides identical to human are highlighted in yellow. Bold capital letters are nucleotides for the TATA box. The region between the two red lines in each alignment is the 24 bp with the highest identity. Identity, shown in the last column, is calculated by dividing the number of nucleotides identical to human in each species by the length of the region (24 bp).

Species	TATA box			Alignment Surrounding TATA box	Identity
	Str	Location	Sequence		
Human	-	chr1:225969260-225969265	TATAAA	NA	NA
Mouse	-	chr8:123505304-123505309	TATAAA	cggaggggaatgcccgcgggctaTATAAAA-cctgagcagaggggacaagcgg gtgggggagagctcccgggactaTATAAAAaccctgtgcaaggggacaggc-g	0.88
Rat	-	chr19:52979006-52979011	TATAAA	cggaggggaatgcccgcgggctaTATAAAA-cctgagcagaggggacaagcgg tgtggagagagctccaggactaTATAAAAaccctgagc-taggggacagggcgg	0.92
Fugu	+	U36850 704-709	TATAAA	c---------ggaggggaatgccccgcgggctaTATAAAAccctgagcagaggcgacaa-agcgg ttttgtgaaatggtggaggtaggg----------TATAAAAgaatgaac-cagcaacacaaaagga	0.71

Table 16. Comparison of regions surrounding the Inr of the transcription factor AP-2 gamma (*TFAP2C*) gene in human, mouse and rat. No *Fugu* sequence is available. Str = Strand. The alignment between human and mouse, rat, and pufferfish is shown. Nucleotides identical to human are highlighted in yellow. Bold capital letters are nucleotides for the Inr. The region between two red lines in each alignment is the 24 bp with the highest identity. Identity, shown in the last column, is calculated by dividing the number of nucleotides identical to human in each species by the length of the region (24 bp).

Species	Inr			Alignment Surrounding Inr	Identity
	Str	Location	Sequence		
Human	+	chr20:55842789-55842795	ACACTGT	NA	NA
Mouse	+	chr2:174482494-174482500	ACACTGT	g t c c c c c g g c t a t c g c c a g g a c A C A C T G T t c g g g c g c g g c t t t c c c c g t c g t c c c c c g g c t a t c g c c g g g a c A C A C T G T g a g g g t g a g g c t t c t c c g g t c	0.96
Rat	+	chr3:137980971-137980977	ACACTGT	g t c c c c c g g c t a t c g c c a g g a c A C A C T G T t c g g g c g c g g c t t t c c c c g t c g t c c c c c g g c t a t c g c c t g g a c A C A C T G T g a g g g t g a g g c t t c t c c g g t c	0.96

on the *Fugu TFAP2C* gene. The *Fugu* homologue was identified by a BLAST search of the *Fugu* genome using *TFAP2C* mRNA from human (UCSC Accession Number: BC035664) and mouse (UCSC Accession Number: X94694). The alignment with the lowest expect value was $5x10^{-5}$ between human and *Fugu* CAAB1001040 and $9x10^{-10}$ between mouse and SCAFFOLD CAAB1001040. As both searches associated SCAFFOLD CAAB1001040 with *TFAP2C* mRNA, this region is likely to contain the *Fugu TFAP2C* gene. However, no alignment for the beginning of either the human or mouse *TFAP2C* mRNA was found within the *Fugu* SCAFFOLD CAAB1001040 sequence. Therefore, another search of the *Fugu* genome was done with the first exon of the human *TFAP2C* gene. Only hits with high expect values (greater than 0.064) were found. Because the initiator element lies upstream of exon 1 and the location *Fugu TFAP2C* exon 1 is unclear, no further analyses of *Fugu* were done. As shown in Table 16, the region upstream and including the Inr box exhibited a 96% identity between the human sequence and those of both the mouse and rat. According to UCSC, there is a 6509 bp CpG island that begins 4118 bp upstream of the TSS in human, a 2264 bp CpG island that begins 44 bp upstream of the TSS in mouse. Rat *TFAP2C* has a 2577 bp CpG island, but because the TSS is not defined in UCSC, its position relative to TSS is uncertain.

The Inr and surrounding region of the *TAF7* gene in human, mouse and rat are summarized in Table 17. The Inr in human TAF7 was described by Zhou *et al.* (2001). The same Inr sequence was found in the *TAF7* gene in mouse and rat by manual inspection. There has been no published research on *Fugu TAF7*. In order to find the pufferfish *TAF7* gene, a BLAST search of the *Fugu* genome was done with the human *TAF7* mRNA sequence (NCBI Accession Number: NM_005642). The alignment with the lowest expect value was 0.16. This expect value does not provide a convincing location for the *TAF7* gene in *Fugu,* so *Fugu* was excluded from the

Table 17. Comparison of regions surrounding the Inr of the TATA box-binding protein-associated factor (*TAF7*) gene in human, mouse and rat. No *Fugu* sequence is available. Str = Strand. Nucleotides identical to human are highlighted in yellow. Bold capital letters are nucleotides for the Inr. The region between the two red lines in each alignment is the 24 bp with the highest identity. Identity, shown in the last column, is calculated by dividing the number of nucleotides identical to human in each species by the length of the region (24 bp).

Species	Inr			Alignment Surrounding Inr	Identity
	Str	Location	Sequence		
Human	-	chr5:140683497-140683503	AGCACTT	NA	NA
Mouse	-	chr18:37918530-37918536	AGCACTT	c t c c c g c c c c g c g t c g c g a c g t **A G C A C T T** c c g t t t t t g c t g g g t a g g c g a g c g t c t g c t t c c g c t c t g a c g g **A G C A C T T** t c g t t t t c t t c a g t g a g a g g a	0.75
Rat	+	chr18:53496673-53496679	AGCACTT	c t c c c g c c c c g c g t c g c g a c g t **A G C A C T T** c c g t t - - - - - t t t g c t g g g t a g g c g a g t c t g c t t c c g c t c a c t g c c g t **A G C A C T T** t c t t t c c t t c t t t t c t g c g t - - - - - g	0.83

analysis of the *TAF7* core promoter. As shown in Table 17, the region surrounding the Inr element exhibits moderate homology between the human, mouse and rat *TAF7* genes. The idenity was 0.75 between human and mouse and 0.83 between human and rat.

The human *TAF7* gene is also associated with DPE, a downstream promoter element (Zhou *et al.* 2001). The location of DPE is chr5:140683467-140683472 on the reverse strand. This location is 29 to 33 bp downstream of the transcription initiation site (which is numbered as +1). DPE is usually found at +28 to +32 bp (Kadonaga, 2002). No homologous DPE sequence could be found between base pairs +28 to +33 in mouse or rat. Because DPE was only found in association with the human *TAF7* locus, no alignment around this core promoter element was done. According to UCSC, there was a 596 bp CpG island starting 196 bp upstream of the TSS in human, and a 390 bp CpG island starting 163 bp upstream of the TSS in mouse. A 353 bp CpG island was present in rat, but because the TSS is not defined, their relative positions could not be determined.

The *LCK* gene in human, mouse, rat, and *Fugu* has two promoters, but no TATA or CAAT box has been found (Brenner *et al.*, 2002). However, a 11 bp PRE (Putative Regulatory Element) sequence that lies 366-376 bp upstream of the *Fugu LCK* gene TSS (NCBI Assession Number AF411956) was found to be identical or highly homologous to regions of both the distal and proximal *LCK* promoter in the mammals (Brenner *et al.*, 2002). The sequences surrounding PRE in its distal and proximal locations are summarized in Tables 18 and 19. The sequence including the distal PRE and slightly downstream exhibited 96% identity between human and mouse, 96% identity between human and rat, and 83% identity between human and *Fugu.* The identity was substantially lower for the PRE associated with the proximal *LCK* promoter.

Table 18. Comparison of regions surrounding the highly-conserved PRE sequence in the distal promoter region of the lymphocyte-specific protein-tyrosine kinase (*LCK*) gene in human, mouse, rat and *Fugu*. Str = Strand. The *Fugu* sequence was obtained from Brenner *et al.*, 2002. Nucleotides identical to human are highlighted in yellow. Bold capital letters are nucleotides for the PRE. The region between the two red lines in each alignment is the 24 bp with the highest identity. Identity, shown in the last column, is calculated by dividing the number of nucleotides identical to human in each species by the length of the region (24 bp).

Species	Distal PRE			Alignment Surrounding Distal PRE	Identity
	Str	Location	Sequence		
Human	+	chr1:32143474-32143484	GGAGGCAGGAA	NA	NA
Mouse	-	chr4:127665040-127665050	GGAGACAGGAA	---ttgggctgcggagccccctccgggaGGAGGGCAGGAAgtcagggtgggacgtgggc ggttgggctgcagagccctcaGGAGA---CAGGAAgtcagggtgggacgtgggc	0.96
Rat	-	chr5:28766131-28766141	GGAGACAGGAA	tgggctgcggagccccctccgggaGGAGGGCAGGAAgtcagggtgggacgtgggc tgggctgcggagccccctcagaaaGGAGACAGGAAgtcagggtgggacgtgggc	0.96
Fugu	+	AF411956:17204-17214	GGAGGCAGGAA	tgggctgcggagccccctccgggaGGAGGGCAGGAAg-------tcagggtgggacgtggc gaccttttgacattatttactttGGAGGGCAGGAAGcccaccccatcagggcttggga------	0.83

Table 19. Comparison of regions surrounding the highly-conserved PRE sequence in the proximal promoter region of the lymphocyte-specific protein-tyrosine kinase (*LCK*) gene in human, mouse, rat, and *Fugu*. Str = Strand. The *Fugu* sequence was obtained from Brenner *et al.*, 2002. Nucleotides identical to human are highlighted in yellow. Bold capital letters are nucleotides for the PRE. The region between the two red lines in each alignment is the 24 bp with the highest identity. Identity, shown in the last column, is calculated by dividing the number of nucleotides identical to human in each species by the length of the region (24 bp).

Species	Proximal PRE			Alignment Surrounding Proximal PRE	Identity
	Str	Location	Sequence		
Human	+	chr1:32166263-32166273	AGAGGCAGGAA	NA	NA
Mouse	-	chr4:127644709-127644719	GGAGGCAAGAA	cagagggagaggggagggggcA----------GAGGCAGGAAgtgggtaactagactaaca -------------gattgggggctctgtcttaactGGAGGCAGGAAgcctctggatcctctgctc	0.71
Rat	-	chr5:28761107-28761115	GAGGCAGGA	cagagggagaggggagggggcAGAGGCAGGAAgtgggtaactagactaaca acccccagcttgttcaagagacaGAGGCAGGAAgaatggaagggttttaggaccct	0.67
Fugu	NA	AF411956 17204-17214	GGAGGCAGGAA	----------cagagggagaggggagggggcAGAGGCAGGAAgtgggtaactagactaaca gaccttttgacattatttactttG----------GAGGCAGGAAaccaccccatcagggcttggga	0.54

According to UCSC, there was no immediate CpG island upstream of the distal promoter region, but a 561 bp CpG island starting 1092 upstream of the proximal TSS in human. Because the Lck gene was not annotated in UCSC for mouse or rat, how the CpG island is distributed with respect to the gene could not be determined.

Overall, the analysis of these five genes shows that core promoter elements can be embedded in larger regions of conserved sequence. Some of the regions of identity observed in association with the core promoters of these other genes were as long and as strong as the one observed with NF1HCS.

3.10 Analyses of the EI1

Exon 1 was used as an anchor for the mVista alignment of Intron 1 but was not considered in the analysis. Using mVista and Frameslider, 4 homologous regions were found in intron 1 in the comparisons of human vs. mouse, and human vs. rat. 175 homologous regions were found in the mouse vs. rat comparison. Only three of the four homologous regions in HvsM were also identical in HvsR. However, comparison with the MvsR alignment reveals an apparent misalignment in HvsR, which leads to the loss of one homology region (EI1-HHR3). If this misalignment is taken into account, HvsR has 5 homology regions, and all 4 regions found in HvsM are found also in human vs. rat (Figure 35).

3.10.1 EI1-HHR1 - Section 1a

EI1-HHR1 is located at Hchr17:29272543-29272633, Mchr11:80125572-80125667, and Rchr10:61761246-61761340 (Figure 36). According to Frameslider, the identity is 0.90-0.92 in

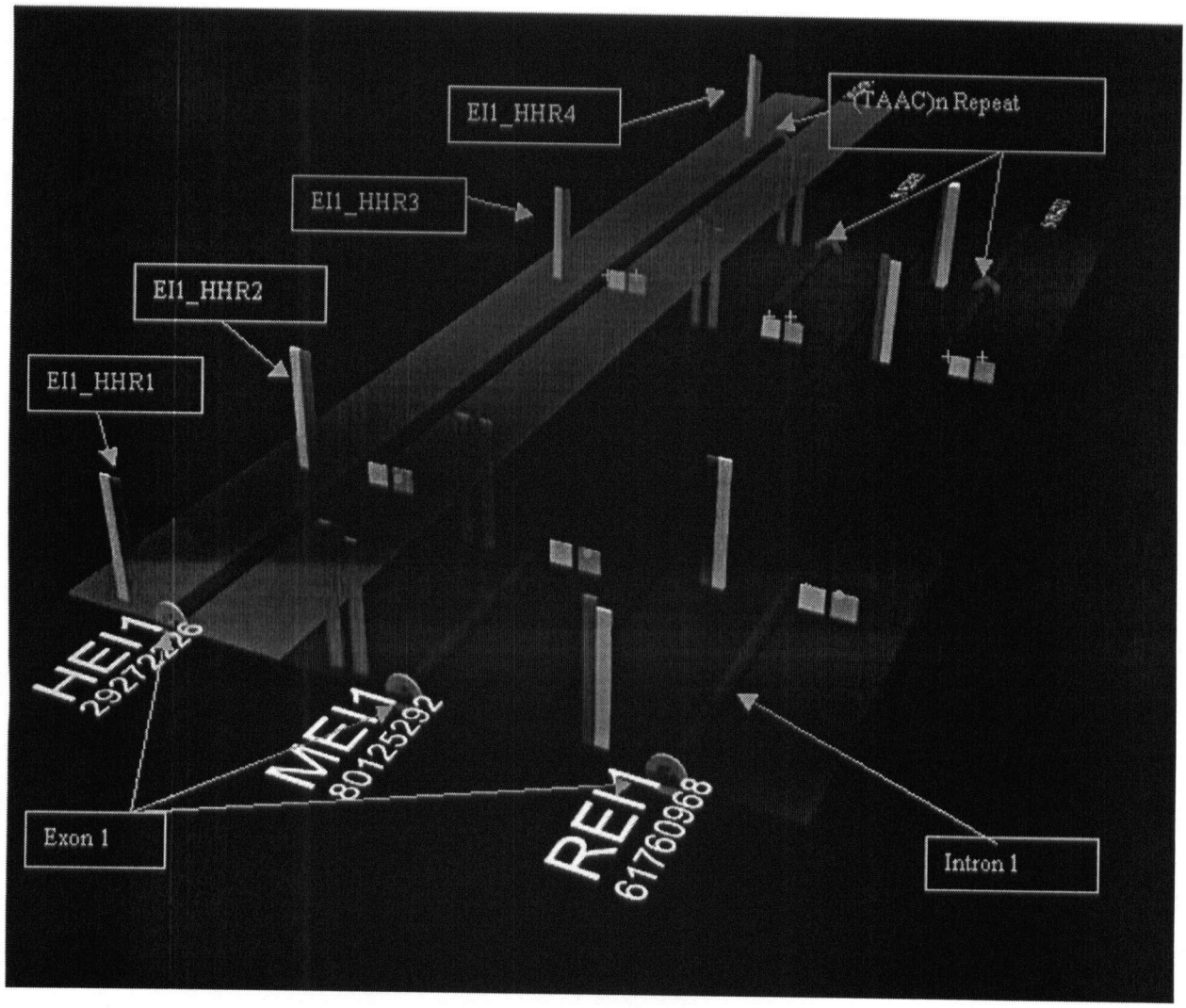


Figure 35. Summary of EI1. Blue rods denote non-coding regions; purple semicircles denote exon 1. Bars perpendicular to the plane are homologies denoted by the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. These bars also indicate the locations of the highly homologous sequences. Tfactor predictions are indicated by boxes under non-coding regions with the follow colours: Yellow – MATCH^{TM}, Orange – MatInspector. Locations of some tfactor predictions are shown. Tfactor binding sites are represented by the following symbols: + for MyoD, • for Pax-2, and * for Pax-4. Note that mouse and rat Intron is shorter than human Intron 1

a) mVista

```
HvsM
Human    accctccatccccttttatcccagcccttccgcttggaaatgg-----ggatgagtgacct
         ||||||||||||||||||||||||||||||||||||||||     |||      ||||||||||||
Mouse    -ccctccatccccttttatcccagcccttccgcttgctcttggtgcggggatgagtgacct

Human    gggggcgcctttaggggcgcgccatctggatttaat
         |||||||| ||||||| |   ||||||||||||||
Mouse    gggggcgctttcagggccactccatctggatttaat

HvsR
Human    accctccatccccttttatcccagcccttccgcttggaaatgg-----ggatgagtgacct
         ||||||||||||||||||||||||||||||||||||||||     |||      ||||||||||||
Rat      -ccctccatccccttttatcccagcccttccgcttgctcttggtgcgaggatgagtgacct

Human    gggggcgcctttaggggcgcgccatctggatttaat
         |||||||| ||||||| |   ||||||||||||||
Rat      gggggcgctttcaggg-cactccatctggatttaat

MvsR
Mouse    ccctccatccccttttatcccagcccttccgcttgctcttggtgcggggatgagtgacctg
         |||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Rat      ccctccatccccttttatcccagcccttccgcttgctcttggtgcgaggatgagtgacctg

Mouse    ggggcgctttcagggccactccatctggatttaat
         ||||||||||||||| ||||||||||||||||||
Rat      ggggcgctttcaggg-cactccatctggatttaat
```

b) Combined Alignment

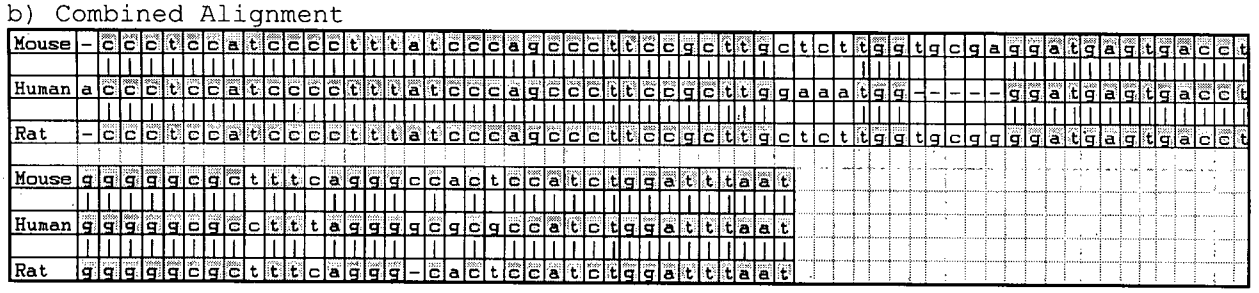


Figure 36. Alignment of EI1-HHR1 in Hchr17:29272543-29272633, Mchr11:80125572-80125667, and Rchr10:61761246-61761340 from (a) mVista and (b) when combined. Highlighted yellow boxes are nucleotides shared by all three species.

HvsM, 0.88-0.92 in HvsR, and 0.98 in MvsR. There is no prediction from GenScan or RepeatMasker in this region for the human sequence. EI1-HHR1 is located 316-436 bp downstream from the translation start site in human, 279-374 bp downstream in mouse, and 277-371 bp downstream in rat.

3.10.2 EI1-HHR1 - Section 1b

Human EI1-HHR1 was compared to the whole *Fugu* EI1 using Pairwise BLAST. There were no hits under the default settings. The parameters were adjusted to search for the longest exact match. At the lowest Wordsize setting, which is 7, there were 27 hits. The longest match that was shared by all four species is 9 bp long. The positions were FCAAB01003481:23273-23281, Hchr17:29272595-29272603, Mchr11:80125629-80125637, and Rchr10:61761303-61761311. Relative to the translation start site, this match begins 544 bp downstream in *Fugu*, but only 370 bp downstream in the human. Considering that *Fugu* intron 1 is 22 times shorter than intron 1 in the mammalian species, one might argue that it is unlikely that this site is further from the translation start site in *Fugu* than in human.

3.10.3 EI1-HHR1 - Section 2

A segment extending from 100 bp downstream and 100 bp upstream of EI1-HHR1 in human, mouse, or rat was downloaded as follows: Hchr17:29272443-29272733, Mchr11:80125472-80125767, and Rchr10:61761146-61761440. These sequences, together with whole *Fugu* EI1, were analyzed using MATCH^{TM} and MatInspector (Table 20 and Table 21).

There were no predictions from MATCH^{TM} under the minFP setting for human, mouse, and rat, but some predictions for *Fugu* EI1. There were several predictions that were shared by the three

Table 20. Summary of MATCH™ predictions surrounding EI1-HHR1 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, and chr10 for rat. Core S. = Core similarity. Matrix S. = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote tfactor binding sites that are detected with both minFN and minFP settings. Bold characters denote the regions within EI1-HHR1. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Human				Mousr				Rat			
	Beginning	End	C.	M.	Beginning	End	C.	M.	Beginning	End	C.	M.
ER	29272519	29272537	1	0.9	80125555	80125573	1	0.9	61761229	61761247	1	0.9
AP-1	29272528	29272538	0.94	0.85	80125564	80125574	0.94	0.86	61761238	61761248	0.94	0.86
c-Rel	**29272563**	**29272572**	**0.95**	**0.77**	**80125592**	**80125601**	**0.95**	**0.77**	**61761266**	**61761275**	**0.95**	**0.77**
Pax-6	**29272564**	**29272584**	**0.79**	**0.62**	**80125593**	**80125613**	**0.79**	**0.65**	**61761267**	**61761287**	**0.79**	**0.65**
myogenin/ NF-1	**29272569**	**29272597**	**0.76**	**0.63**	**80125598**	**80125626**	**0.93**	**0.66**	**61761272**	**61761300**	**0.93**	**0.64**
ER	**29272581**	**29272599**	**1**	**0.93**	**80125615**	**80125633**	**1**	**0.94**	**61761289**	**61761307**	**1**	**0.93**
COMP1	**29272583**	**29272606**	**0.79**	**0.7**	**80125617**	**80125640**	**0.79**	**0.7**	**61761291**	**61761314**	**0.79**	**0.69**
AP-1	**29272590**	**29272600**	**0.94**	**0.87**	**80125624**	**80125634**	**0.94**	**0.87**	**61761298**	**61761308**	**0.94**	**0.87**
Pax-4	**29272603**	**29272623**	**0.79**	**0.74**	**80125637**	**80125657**	**0.98**	**0.71**	**61761311**	**61761331**	**0.98**	**0.7**
Hand1/E47	***29272615***	***29272630***	***1***	***0.93***	***80125649***	***80125664***	***1***	***0.93***	***61761322***	***61761337***	***1***	***0.93***
COMP1	**29272620**	**29272643**	**0.79**	**0.6**	**80125654**	**80125677**	**0.79**	**0.57**	**61761327**	**61761350**	**0.79**	**0.56**
HNF-1	29272627	29272641	1	0.85	80125661	80125675	1	0.77	61761334	61761348	1	0.78
Pax-4	29272672	29272692	0.79	0.69					61761382	61761402	0.79	0.71
Pax-6	29272672	29272692	0.78	0.63					61761382	61761402	0.78	0.56
v-Myb	29272713	29272722	0.94	0.89	80125746	80125755	0.94	0.89	61761423	61761432	0.94	0.91

Table 21. Summary of MatInspector predictions surrounding EI1-HHR1 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, and chr10 for rat. C = Core similarity. M = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote predictions that are identified when the core similarity setting is 1.00. Bold characters denote the regions within EI1-HHR1. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Further Information	Human				Mouse				Rat			
		Beginning	End	Core S	Matrix S.	Beginning	End	Core S.	Matrix S	Beginning	End	Core S	Matrix S.
V$MAZF/ MAZR.01	*MYC-associated zinc finger protein related transcription factor*	*29272511*	*29272523*	*1*	*0.939*	*80125547*	*80125559*	*1*	*0.939*	*61761221*	*61761233*	*1*	*0.939*
V$SP1F/ GC.01	GC box elements	29272510	29272524	0.872	0.951	80125546	80125560	0.872	0.951	61761220	61761234	0.872	0.951
V$SP1F/ SP1.01	stimulating protein 1 SP1, ubiquitous zinc finger transcription factor	29272510	29272524	0.819	0.935	80125546	80125560	0.819	0.935	61761220	61761234	0.819	0.935
V$MOKF/ MOK2.02	*Ribonucleoprotein associated zinc finger protein MOK-2 (human)*	29272508	29272528	*1*	0.996	*80125544*	*80125564*	*1*	0.996	61761218	61761238	*1*	0.996
V$PBXC/ PBX1.ME	***Binding site for a Pbx1/Meis1 heterodimer***	***29272584***	***29272600***	***1***	***0.798***	***80125618***	***80125634***	***1***	***0.798***	***61761292***	***61761308***	***1***	***0.798***
V$PBXC/ PBX1.ME	**Binding site for a Pbx1/Meis1 heterodimer**	**29272584**	**29272600**	**0.757**	**0.746**	**80125618**	**80125634**	**0.757**	**0.746**	**61761292**	**61761308**	**0.757**	**0.746**
V$PBXC/ PBX1.ME	**Binding site for a Pbx1/Meis1 heterodimer**	**29272584**	**29272600**	**0.75**	**0.78**	**80125618**	**80125634**	**0.75**	**0.78**	**61761292**	**61761308**	**0.75**	**0.78**
V$RXRF/ FXRE.01	***Farnesoid X - activated receptor (RXR/FXR dimer)***	***29272585***	***29272601***	***1***	***0.887***	***80125619***	***80125635***	***1***	***0.887***	***61761293***	***61761309***	***1***	***0.887***
V$ZF5F/ ZF5.01	***Zinc finger / POZ domain transcription factor***	***29272598***	***29272608***	***1***	***0.954***	***80125632***	***80125642***	***1***	***0.951***	***61761306***	***61761316***	***1***	***0.951***
V$NEUR/ NEUROD1	***DNA binding site for NEUROD1 (BETA-2 / E47 dimer)***	***29272616***	***29272628***	***1***	***0.899***	***80125650***	***80125662***	***1***	***0.836***	***61761323***	***61761335***	***1***	***0.836***
V$PDX1/ PDX1.01	***Pdx1 (IDX1/IPF1) pancreatic and intestinal homeodomain TF***	***29272624***	***29272644***	***1***	***0.812***	***80125655***	***80125675***	***1***	***0.822***				

mammalian species under the minFN setting. However, there was no prediction made in *Fugu* that was common to the other three species at the location where the match was predicted in the previous section.

Although there were also several predictions for tfactor binding sites from MatInspector, none of these predictions was shared between MatInspector and MATCH^{TM}. None of the predictions found in *Fugu* at the 9 bp exact match was found in mammalian species (not shown). Also, the shared prediction at the matched region for human and mouse was Pdx1, and it was not shared by rat.

The location of EI1-HHR1 and its surrounding tfactor predictions are summarized by Sockeye in Figure 37.

3.10.4 EI1-HHR2 – Section 1a

EI1-HHR2 is located at Hchr17:29281291-29281430, Mchr11:80132603-80132753, and Rchr10:61769227-61769379 (Figure 38). According to Frameslider, the identity is 0.90-0.96 in HvsM, 0.88-0.96 in HvsR, and 0.94-1.00 in MvsR. There is no prediction from GenScan other than *NF1*, and no repeats were detected by RepeatMasker in this region for the human. The downstream locations relative to the translation start sites are 9066-9203 bp in human, 7312-7462 bp in mouse, and 8260-8412 bp in rat.

3.10.5 EI1-HHR2 – Section 1b

When human EI1-HHR2 was compared to *Fugu* EI1 under the default settings, no region of high homology is found. There were 19 exact matches of 7 bp long. There was only 1 exact match between human and *Fugu* that is 8 bp long, and this match is not shared with mouse and rat.

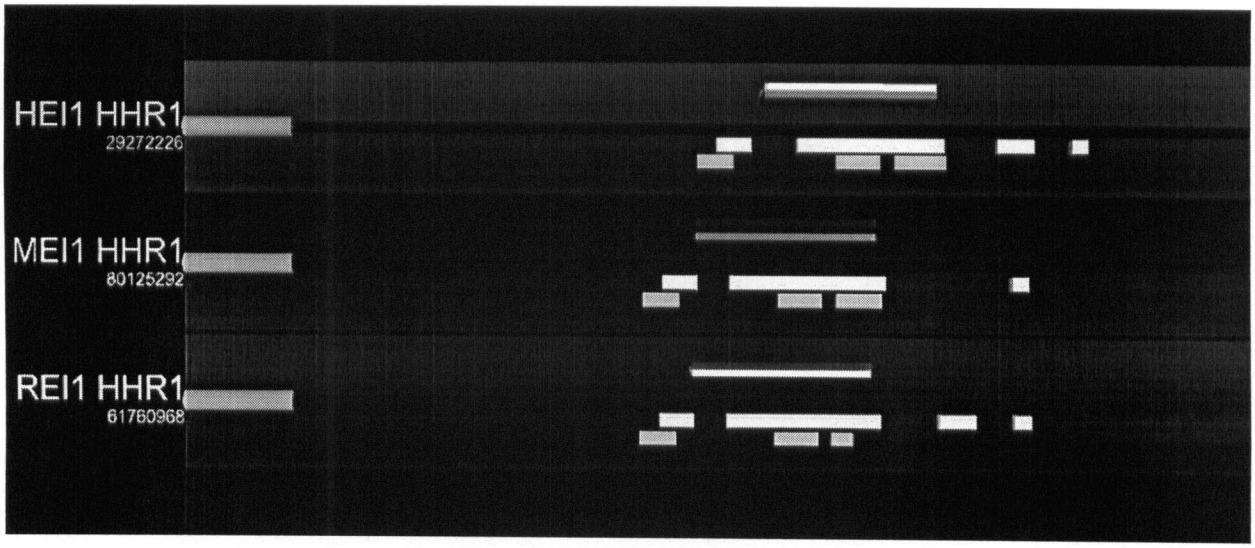


Figure 37. Sockeye presentation of EI1-HHR1 and related tfactor predictions at Hchr17:29272443-29272733, Mchr11:80125472-80125767, and Rchr10:61761146-61761440. Blue rods denote introns; purple rods denote exons. Bars above the introns represent the sequences for EI1-HHR1 with homologies denoted by the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. Boxes below are tfactor predictions from MATCH^{TM} (yellow) and MatInspector (orange). Precise locations of tfactor predictions can be found in Tables 20 and 21.

a) mVista

```
HvsM
Human     aggaggtatcggaaggctct-------aaaggaagtttaaggaggagaatattctata--
            ||||||| ||||||||||        || ||||||||||||||||||| |||||||
Mouse     gagaggtataggaaggctcttgttttaaatggaagtttaaggaggagaatgttctataaa

Human     -gaagtggaaggggagatttgtggtcagcttaaactgttaaaaggcttgggatcaatact
            |||||||||||||||  |||||||||||||||||||||||||||||| ||||||||||
Mouse     agaagtggaaggggaggattgtggtcagcttaaactgttaaaaggcttaggatcaatact

Human     gaagcagaatatgagcatcttaatctgt
            ||||||||||||||||||||||| ||||
Mouse     gaagtagaatatgagcatcttaagctat

HvsR
Human     aggaggtatcggaaggctct----------aaaggaagtttaaggaggagaatattctat
            ||||||| ||||||||||          || |||||||||||||||||| ||||||
Rat       gagaggtataggaaggctctttttttttaaacggaagtttaaggaggagaatgttctat

Human     a---gaagtggaaggggagatttgtggtcagcttaaactgttaaaaggcttgggatcaat
          |   |||||||||||||||  ||||||||||||||||||||||||||||| ||||||||
Rat       aaaagaagtggaaggggaggattgtggtcagcttaaactgttaaaaggcttaggatcaat

Human     accgaagcagaatatgagcatcttaatctgt
          || |||| |||||||||||||||||| || |
Rat       actgaagtagaatatgagcatcttaagctat

MvsR
Mouse     gagaggtataggaaggctcttgtttt---aaatggaagtttaaggaggagaatgttctat
          |||||||||||||||||||||| ||||   ||| ||||||||||||||||||||||||||
Rat       gagaggtataggaaggctctttttttttaaacggaagtttaaggaggagaatgttctat

Mouse     aaaagaagtggaaggggaggattgtggtcagcttaaactgttaaaaggcttaggatcaat
          ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Rat       aaaagaagtggaaggggaggattgtggtcagcttaaactgttaaaaggcttaggatcaat

Mouse     actgaagtagaatatgagcatcttaagctat
          ||||||||||||||||||||||||||||||
Rat       actgaagtagaatatgagcatcttaagctat
```

b) Combined alignment

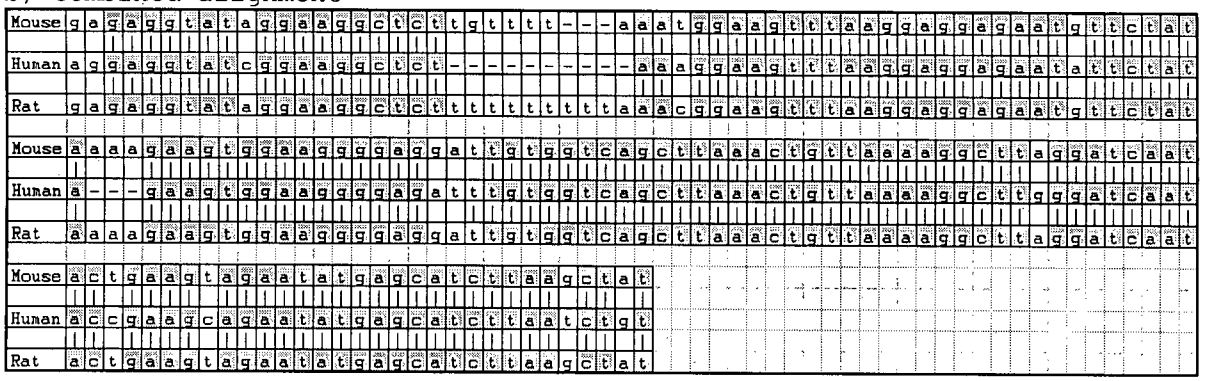


Figure 38. Alignment of EI1-HHR2 at Hchr17:29272443-29272733, Mchr11:80125472-80125767, and Rchr10:61761146-61761440 from (a) mVista and (b) when combined. Nucleotides highlighted in yellow are shared by all three species.

This is a rather unusual result because most other highly homologous regions have at least one 8 bp exact match shared by the four species. EI1-HHR2, which at about 140 bp is longer than any other highly homologous region found in this study except 5UR-HHR3, does not have such a match.

3.10.6 EI1-HHR2 – Section 2

The region extending from 100 bp downstream to 100 bp upstream of the EI1-HHR2 in human, mouse, and rat were downloaded as follows: Hchr17:29281191-29281530, Mchr11:80132503-80132853, and Rchr10:61769127-61769479. These sequences were analyzed using MATCH™ and MatInspector (Table 22 and Table 23). Since there was no convincing match from *Fugu*, special attention was focused on predictions within the EI1-HHR2 region as a whole.

There was no prediction from MATCH™ under the minFP setting that corresponded in human and mouse. Although rat has two predictions under the minFP setting, neither prediction is shared by all three species. There was a long list of predictions that were shared by all three species under the minFN setting.

The list of predictions from Matinspector was shorter and most of the predictions had a core similarity of 1. The prediction of Pax-2 from MatInspector at Hchr17:29281355-29281377, Mchr11:80132678-80132700, and Rchr10:61769304-61769326, and the prediction of Pax-4 from MATCH™ at Hchr17:29281360-29281380, Mchr11:80132683-80132703, and Rchr10:61769309-61769329, belonged to the same tfactor family. Furthermore, this particular tfactor is within EI1-HHR2. However, although the two predictions overlapped, the core sequences within the predictions had no overlap. Otherwise, there are no shared predictions between the two programs.

Table 22. Summary of MATCH™ predictions surrounding EI1-HHR2 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, chr10 for rat. Core S. = Core similarity. Matrix S. = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote tfactor binding sites that are detected with both minFN and minFP settings. Bold characters denote the regions within EI1-HHR2. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Human				Mouse				Rat			
	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.
GATA-1	*29281192*	*29281205*	*0.986*	*0.956*	*80132506*	*80132519*	*0.991*	*0.97*				
CDP CR1	29281193	29281202	0.768	0.739	80132507	80132516	0.858	0.805				
Evi-1	29281208	29281222	0.842	0.602					61769138	61769152	0.773	0.694
CDP CR1	29281208	29281217	0.794	0.755	80132524	80132533	0.929	0.65				
c-Rel	*29281221*	*29281230*	*1*	*0.866*	*80132534*	*80132543*	*1*	*0.957*	61769154	61769163	0.756	0.797
Pax-4	29281237	29281257	0.789	0.593	80132547	80132567	0.842	0.72	*61769167*	*61769187*	*1*	*0.773*
Pax-6	29281237	29281257	0.778	0.591	*80132547*	*80132567*	*1*	*0.773*	61769167	61769187	0.842	0.734
HNF-4	29281276	29281294	0.848	0.732	80132588	80132606	0.848	0.596	61769212	61769230	0.848	0.596
Elk-1	**29281286**	**29281301**	**0.9**	**0.827**	**80132592**	**80132607**	**0.938**	**0.862**	**61769216**	**61769231**	**0.938**	**0.862**
Elk-1	***29281295***	***29281308***	***1***	***0.943***	**80132607**	**80132620**	**0.938**	**0.887**	**61769231**	**61769244**	**0.938**	**0.887**
Elk-1	**29281308**	**29281321**	**0.938**	**0.912**	**80132627**	**80132640**	**0.924**	**0.899**	**61769254**	**61769267**	**1**	**0.967**
Elk-1	**29281308**	**29281323**	**1**	**0.819**	**80132627**	**80132642**	**1**	**0.81**	**61769254**	**61769269**	**1**	**0.845**
HNF-4	**29281314**	**29281332**	**0.763**	**0.706**	**80132633**	**80132651**	**0.763**	**0.706**	**61769260**	**61769278**	**0.763**	**0.706**
Pax-6	**29281315**	**29281335**	**0.792**	**0.622**	**80132634**	**80132654**	**0.792**	**0.619**	**61769261**	**61769281**	**0.792**	**0.619**
Pax-4	**29281316**	**29281336**	**0.789**	**0.592**	**80132635**	**80132655**	**0.789**	**0.592**	**61769262**	**61769282**	**0.789**	**0.592**
Oct-1	**29281322**	**29281336**	**0.781**	**0.663**	**80132641**	**80132655**	**0.781**	**0.662**	**61769268**	**61769282**	**0.781**	**0.662**
FOXD3	**29281330**	**29281341**	**0.944**	**0.796**	**80132649**	**80132660**	**0.996**	**0.83**	**61769276**	**61769287**	**0.996**	**0.83**
Elk-1	**29281341**	**29281354**	**0.924**	**0.877**	**80132663**	**80132676**	**0.924**	**0.877**	**61769290**	**61769303**	**0.924**	**0.877**
Nkx2-5	**29281341**	**29281347**	**1**	**0.897**	**80132663**	**80132669**	**1**	**0.897**	**61769290**	**61769296**	**1**	**0.897**
Pax-4	**29281360**	**29281380**	**0.789**	**0.609**	**80132683**	**80132703**	**0.789**	**0.609**	**61769309**	**61769329**	**0.789**	**0.609**
v-Myb	**29281370**	**29281379**	**0.938**	**0.854**	**80132693**	**80132702**	**0.938**	**0.854**	**61769319**	**61769328**	**0.938**	**0.854**
Oct-1	**29281371**	**29281385**	**0.792**	**0.662**	**80132694**	**80132708**	**0.792**	**0.662**	**61769320**	**61769334**	**0.792**	**0.662**
HNF-4	**29281373**	**29281391**	**0.883**	**0.772**	**80132696**	**80132714**	**0.883**	**0.772**	**61769322**	**61769340**	**0.883**	**0.772**
Barbie Box	**29281377**	**29281391**	**0.979**	**0.896**	**80132700**	**80132714**	**0.979**	**0.897**	**61769326**	**61769340**	**0.979**	**0.897**
SOX-9	**29281388**	**29281401**	**0.925**	**0.878**	**80132711**	**80132724**	**0.925**	**0.879**	**61769337**	**61769350**	**0.803**	**0.602**
Pax-4	**29281388**	**29281408**	**0.803**	**0.624**	**80132711**	**80132731**	**0.803**	**0.602**	**61769337**	**61769357**	**0.925**	**0.879**
CDP CR1	**29281391**	**29281400**	**0.865**	**0.824**	**80132714**	**80132723**	**0.865**	**0.811**	**61769340**	**61769349**	**0.865**	**0.811**
ER	29281401	29281419	0.845	0.802	80132724	80132742	0.845	0.805	61769350	61769368	0.845	0.805
COMP1	*29281436*	*29281458*	*1*	*0.79*	*80132753*	*80132776*	*1*	*0.753*	*61769379*	*61769402*	*1*	*0.753*
CDP CR1	29281437	29281446	0.768	0.806	80132755	80132764	0.762	0.801	61769381	61769390	0.762	0.801
CDP CR1	29281441	29281450	0.929	0.675	80132759	80132768	0.929	0.675	61769385	61769394	0.929	0.675
AP-1	29281442	29281452	1	0.932	80132760	80132770	1	0.932	61769386	61769396	1	0.932
AP-1	29281442	29281452	1	0.895	80132760	80132770	1	0.898	61769386	61769396	1	0.898
HFH-3	29281448	29281460	0.981	0.905	80132764	80132776	1	0.916	61769390	61769402	1	0.916
TATA	29281453	29281462	1	0.936	80132769	80132778	1	0.936	61769395	61769404	1	0.936
COMP1	29281457	29281480	0.914	0.712	80132778	80132801	0.786	0.562				
HNF-3beta	29281466	29281480	1	0.879	80132782	80132796	1	0.856	61769406	61769422	1	0.884
FOXD3	29281468	29281479	1	0.855	80132784	80132795	1	0.85	61769410	61769421	1	0.883
Pax-4	29281470	29281490	0.81	0.608	80132786	80132606	0.81	0.596				
AP-1	29281478	29281488	0.935	0.852					61769420	61769430	1	0.909
Elk-1	29281510	29281523	0.927	0.914	80132826	80132839	1	0.937	61769453	61769466	0.927	0.872
Evi-1	29281515	29281529	0.826	0.643	80132826	80132840	1	0.706	61769453	61769467	1	0.717

Table 23. Summary of MatInspector predictions surrounding EI1-HHR2 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, and chr10 for rat. C = Core similarity. M = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote predictions that are identified when the core similarity setting is 1.00. Bold characters denote the regions within EI1-HHR2. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Further information	Human				Mouse				Rat			
		Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.
V$HOMS/ S8.01	*Binding site for S8 type homeodomains*	*29281209*	*29281217*	*1*	*0.997*					*61769141*	*61769149*	*1*	*0.996*
V$NKXH/ DLX1.01	*DLX-1, -2, and -5 binding sites*	*29281208*	*29281220*	*1*	*0.982*					*61769140*	*61769152*	*1*	*0.982*
V$MYT1/ MYT1.02	***MyT1 zinc finger transcription factor involved in primary neurogenesis***	***29281313***	***29281325***	***1***	***0.889***	***80132632***	***80132644***	***1***	***0.887***	***61769259***	***61769271***	***1***	***0.888***
V$HEAT/ HSF1.01	***Heat shock factor 1***	***29281329***	***29281339***	***1***	***0.977***	***80132648***	***80132658***	***1***	***0.987***	***61769275***	***61769285***	***1***	***0.987***
V$NKXH/ HMX3.01	***H6 homeodomain HMX3/Nkx5.1 transcription factor***	***29281339***	***29281351***	***1***	***0.892***	***80132661***	***80132673***	***1***	***0.9***	***61769288***	***61769300***	***1***	***0.9***
V$ETSF/ GABP.01	***GABP: GA binding protein***	***29281341***	***29281357***	***1***	***0.855***					***61769290***	***61769306***	***1***	***0.855***
V$PAX2/ PAX2.01	***Zebrafish PAX2 paired domain protein***	***29281355***	***29281377***	***1***	***0.811***	***80132678***	***80132700***	***1***	***0.886***	***61769304***	***61769326***	***1***	***0.886***
V$EREF/ ER.01	***Estrogen receptor***	***29281360***	***29281378***	***1***	***0.909***	***80132683***	***80132701***	***1***	***0.909***	***61769309***	***61769327***	***1***	***0.909***
V$PBXC/ PBX1_MEIS1.03	*Binding site for a Pbx1/Meis1 heterodimer*	*29281436*	*29281452*	*1*	*0.794*	*80132754*	*80132770*	*1*	*0.785*	*61769380*	*61769396*	*1*	*0.785*
V$PBXC/ PBX1_MEIS1.01	Binding site for a Pbx1/Meis1 heterodimer	29281436	29281452	0.747	0.761	80132754	80132770	0.747	0.754	61769380	61769396	0.747	0.754
V$PBXC/ PBX1_MEIS1.02	Binding site for a Pbx1/Meis1 heterodimer	29281436	29281452	0.75	0.774	80132754	80132770	0.75	0.776	61769380	61769396	0.75	0.776
V$TBPF/ TATA.01	*cellular and viral TATA box elements*	*29281452*	*29281468*	*1*	*0.973*	*80132768*	*80132784*	*1*	*0.979*	*61769394*	*61769410*	*1*	*0.973*
V$TBPF/ TATA.02	*Mammalian C-type LTR TATA box*	*29281452*	*29281468*	*1*	*0.931*	*80132768*	*80132784*	*1*	*0.931*	*61769394*	*61769410*	*1*	*0.931*
V$TBPF/ ATATA.01	Avian C-type LTR TATA box	29281452	29281468	0.75	0.834	80132768	80132784	0.75	0.834	61769394	61769410	0.75	0.834
V$TBPF/ MTATA.01	*Muscle TATA box*	*29281452*	*29281469*	*1*	*0.845*	*80132768*	*80132784*	*1*	*0.874*	*61769394*	*61769410*	*1*	*0.866*

The location of EI1-HHR2 and related tfactor predictions as summarized by Sockeye are shown in Figure 39.

3.10.7 EI1-HHR3 - Section 1a

EI1-HHR3 is located at Hchr17:29299920-29299983, Mchr11:80151206-80151279, and Rchr10:61786728-61786801 (Figure 40). According to Frameslider, the identity is 0.9 in HvsM, 0 in HvsR, and 0.86-0.94 in MvsR. However, the alignment of human and rat provided by mVista was incorrect because of an extensive gap in the alignment. The same region in rat can be aligned to human with an identity of 0.90-0.92. There is no prediction from GenScan or RepeatMasker in this region for the human. The downstream locations relative to the translation start sites are 27695-27758 bp in human, 25915-25988 bp in mouse, 25761-25834 bp in rat.

3.10.8 EI1-HHR3 - Section 1b

Human EI1-HHR3 was compared to the whole *Fugu* EI1 using Pairwise BLAST. There were no hits under the default settings. The parameters were, therefore, adjusted to search for the longest exact match. At the lowest Wordsize setting, which is 7, there were only 2 hits. Neither of these hits was shared among all four species. No hit was detected for an exact match of 8 bp long.

3.10.9 EI1-HHR3 - Section 2

The regions extending from 100 bp downstream to 100 bp upstream of EI1-HHR3 in human, mouse, and rat were downloaded as follows: Hchr17:29299820-29300083, Mchr11:80151106-80151379, and Rchr10:61786628-61786901. These sequences were analyzed using MATCH^{TM} and MatInspector (Table 24 and Table 25). Since there was no convincing match from *Fugu*, special attention was focused on the predictions within the EI1-HHR3 region as a whole.

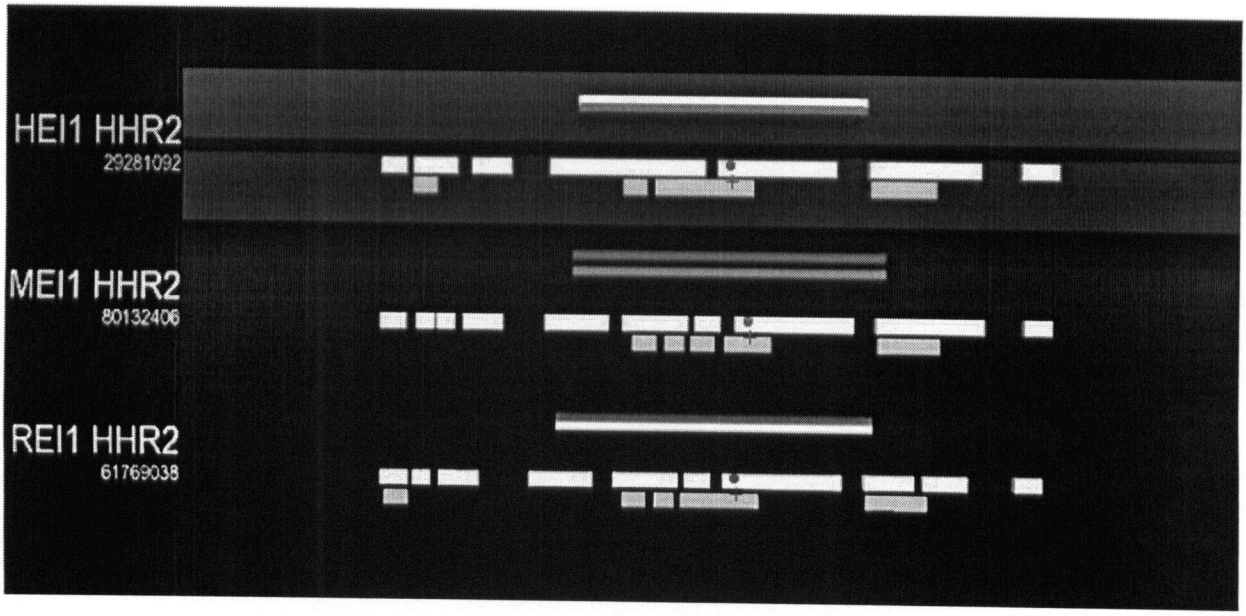


Figure 39. Sockeye presentation of EI1-HHR2 and related tfactor predictions at Hchr17:29281191-29281530, Mchr11:80132503-80132853, and Rchr10:61769127-61769479. Blue rods denote non-coding regions. Bars above the introns represent the sequences for EI1-HHR2 with homologies denoted by the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. Boxes below are tfactor predictions from MATCH^{TM} (yellow) and MatInspector (orange). Locations of Pax-4 (•) from MATCH^{TM} and Pax-2 (+) from MatInspector are shown. Precise locations of other tfactor predictions can be found in Tables 22 and 23.

a) mVista

```
HvsM
Human    aagcttctggcttgaattaagttataaac-ttagcacagtggcaggtgcttgaactgc--
         ||||||||| ||||| |||||||||||  |||||| ||||||||||||| |||||
Mouse    cagcttctggtttgaactaagttataaaaattagcatagtggcaggtgcttcaactgtta

Human    -------catgtta
                 |||||||
Mouse    tatatgccatgtta


MvsR
Mouse    cagcttctggtttgaactaagttataaaaattagcatagtggcaggtgcttcaactgtta
         |||||||||||||||||||||||| ||||| | |||||||||||||||||| ||||||||
Rat      cagcttctggtttgaactaagttgtaaaatt-agcatagtggcaggtgcttaaactgtta

Mouse    -tatatgccatgtta
          || |  |||||| |
Rat      atagacaccatgtca
```

b) mVista with corrected alignment.

```
HvsR
Human    aagcttctggcttgaattaagttataaacttagcacagtggcaggtgcttgaactgc---
         ||||||||| ||||| |||||||||||| |||||| |||||||||||||| |||||
Rat      cagcttctggtttgaactaagttgtaaaattagcatagtggcaggtgcttaaactgttat

Human    ------catgtta
               ||||| |
Rat      agacaccatgtca
```

c) combined alignment

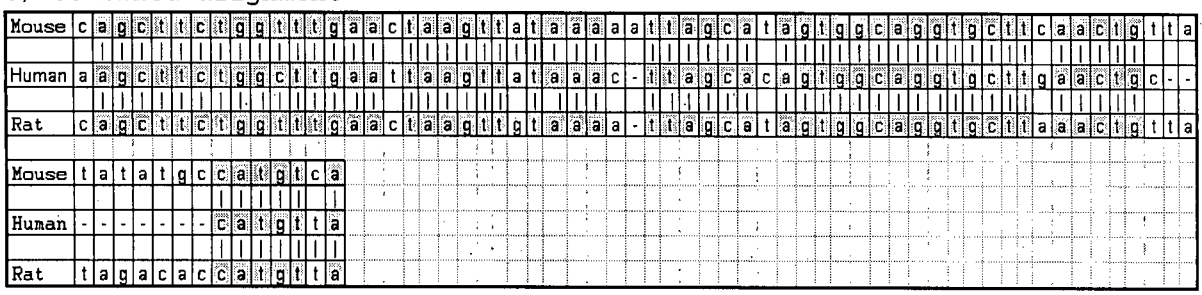


Figure 40. Alignment of EI1-HHR3 in Hchr17:29299920-29299983, Mchr11:80151206-80151279, and Rchr10:61786728-61786801 from mVista (a). (b) Corrected HvsR alignment based on the homology between mouse and rat. (c) Combined alignment, with nucleotides shared by all three species highlighted in yellow.

Table 24. Summary of MATCH^{TM} predictions surrounding EI1-HHR3 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, and chr10 for rat. Core S. = Core similarity. Matrix S. = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote tfactor binding sites that are detected with both minFN and minFP settings. Bold characters denote the regions within EI1-HHR3. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Human				Mouse				Rat			
	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.
FOXD3	29299821	29299832	0.948	0.801	80151110	80151121	0.948	0.809	61786632	61786643	0.948	0.809
Hand1/E47	29299825	29299840	0.871	0.823	80151114	80151129	0.871	0.822				
Pax-4	29299846	29299865	0.81	0.605	80151132	80151151	0.802	0.763	61786654	61786673	0.81	0.615
Pax-6	29299846	29299866	0.772	0.749	80151132	80151152	0.81	0.615	61786654	61786674	0.802	0.763
COMP1	29299849	29299872	0.822	0.581	80151135	80151158	0.822	0.556	61786657	61786680	0.822	0.556
CDP CR1	29299852	29299861	0.775	0.74	80151138	80151147	0.775	0.74	61786660	61786669	0.775	0.74
Pax-4	29299865	29299885	0.817	0.626	80151151	80151171	0.817	0.587	61786673	61786693	0.817	0.59
Pax-4	29299869	29299889	0.789	0.623					61786677	61786697	0.789	0.587
Oct-1	29299875	29299889	0.77	0.701	80151167	80151181	0.893	0.74				
CCAAT box	29299890	29299901	0.953	0.886	80151176	80151187	0.953	0.922	61786698	61786709	0.953	0.886
Hand1/E47	29299918	29299933	1	0.936	80151204	80151219	1	0.922	61786726	61786741	1	0.924
COMP1	**29299923**	**29299946**	**0.786**	**0.551**	**80151209**	**80151232**	**0.822**	**0.569**	**61786731**	**61786754**	**0.822**	**0.575**
HLF	**29299935**	**29299944**	**0.802**	**0.816**	**80151226**	**80151235**	**0.892**	**0.834**	**61786748**	**61786757**	**0.784**	**0.801**
E47	**29299956**	**29299971**	**1**	**0.95**	**80151243**	**80151258**	**1**	**0.95**	**61786764**	**61786779**	**1**	**0.95**
E47	**29299957**	**29299971**	**1**	**0.915**	**80151244**	**80151258**	**1**	**0.915**	**61786765**	**61786779**	**1**	**0.912**
MyoD	**29299958**	**29299969**	**1**	**0.934**	**80151245**	**80151256**	**1**	**0.934**	**61786766**	**61786777**	**1**	**0.934**
v-Myb	**29299968**	**29299977**	**0.938**	**0.852**	**80151255**	**80151264**	**0.938**	**0.868**	**61786776**	**61786785**	**0.938**	**0.854**
Oct-1	**29299974**	**29299988**	**0.792**	**0.753**	**80151265**	**80151279**	**0.946**	**0.882**	**61786777**	**61786791**	**0.792**	**0.721**
Pax-4	29300029	29300049	0.888	0.601	80151326	80151346	0.888	0.625				
Oct-1	29300049	29300063	0.776	0.806	80151346	80151360	0.781	0.807				
Pax-4	29300058	29300078	0.794	0.591	80151355	80151375	0.794	0.613				

Table 25. Summary of MatInspector predictions surrounding EI1-HHR3 on the same strand. 'Beginning' and 'End' represent the corresponding positions on chromosome 17 for human, chr11 for mouse, and chr10 for rat. C = Core similarity. M = Matrix similarity. Only tfactor binding site predictions shared between human and at least one other species at the aligned positions are shown. Italic characters denote predictions that are identified when the core similarity setting is 1.00. Bold characters denote the regions within EI1-HHR3. Boxes highlighted in yellow are tfactor predictions shared by 2 species; orange, 3.

Tfactor	Transcription Factor	Human				Mouse				Rat			
		Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.	Beginning	End	Core S.	Matrix S.
V$MYOD/ MYOD.01	Myoblast determination gene product	*29299957*	*29299971*	*1*	*0.915*	*80151244*	*80151258*	*1*	*0.915*	*61786765*	*61786779*	*1*	*0.986*
V$HEAT/ HSF1.01	Heat shock factor 1	29300056	29300066	0.916	0.937	*80151353*	80151363	*1*	*0.957*	*61786881*	*61786891*	*1*	*0.957*

There was no prediction from MATCH^{TM} under the minFP setting in any of the three species. Among the many predictions shared between the three species under the minFN setting, there were seven conserved predictions within the highly homologous region.

The list of predictions from Matinspector has only two shared predictions, and only one is within the highly homologous region. However, this prediction is also shared by MATCH^{TM} in the same location. At Hchr17:29299957-29299971, Mchr11:80151244-80151258, and Rchr10:61786765-61786779, both programs predict a binding site for MyoD (Myoblast determination gene product), which is involved in myogenic differentiation and inhibition of cell proliferation.

The location of EI1-HHR3 and related tfactor predictions are summarized by Sockeye in Figure 41.

3.10.10 EI1-HHR4 - Section 1a

EI1-HHR4 is the last highly homologous region that was found in EI1. It is located at Hchr17:29323319-29323389, Mchr11:80161561-80161634, and Rchr10:61795122-61795754. According to Frameslider, the identity is 0.90-0.94 in HvsM, 0.88-1.00 in HvsR, and 0.92-0.94 in MvsR. The downstream locations relative to the translation start sites are 51094-51164 bp in human, 36270-36343 bp in mouse, 34155-34787 bp in rat. Inspection of this alignment showed two problems (Figure 42). First, in HvsR, the homology is an artifact because Frameslider is not sensitive to gaps in the primary strand. Second, even though the alignments for HvsM and MvsR are correct, they lie within a repeat region as confirmed by RepeatMasker. Regions Hchr17:29323326-29323377, Mchr11:80161568-80161619, and Rchr10:61795703-61795742 consist of (TCCA)n simple repeats. Information on this region is summarized by Sockeye in

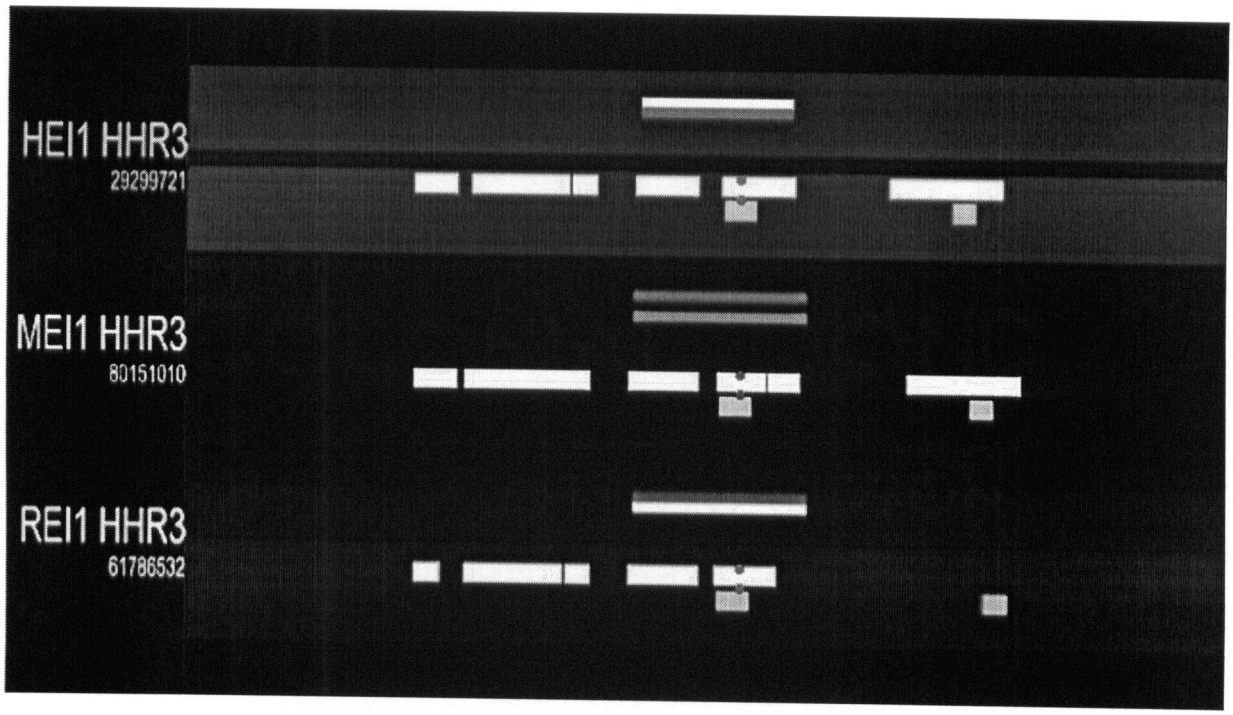


Figure 41. Sockeye presentation of EI1-HHR3 and related tfactor predictions at Hchr17:29299820-29300083, Mchr11:80151106-80151379, and Rchr10:61786628-61786901. Blue rods denote introns. Bars above the introns represent sequences of EI1-HHR3 with homologies denoted by the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. Boxes below are tfactor predictions from MATCH^{TM} (yellow) and MatInspector (orange). Locations of MyoD (•) from MATCH^{TM} and MatInspector are shown. Precise locations of other tfactor predictions can be found in Tables 24 and 25.

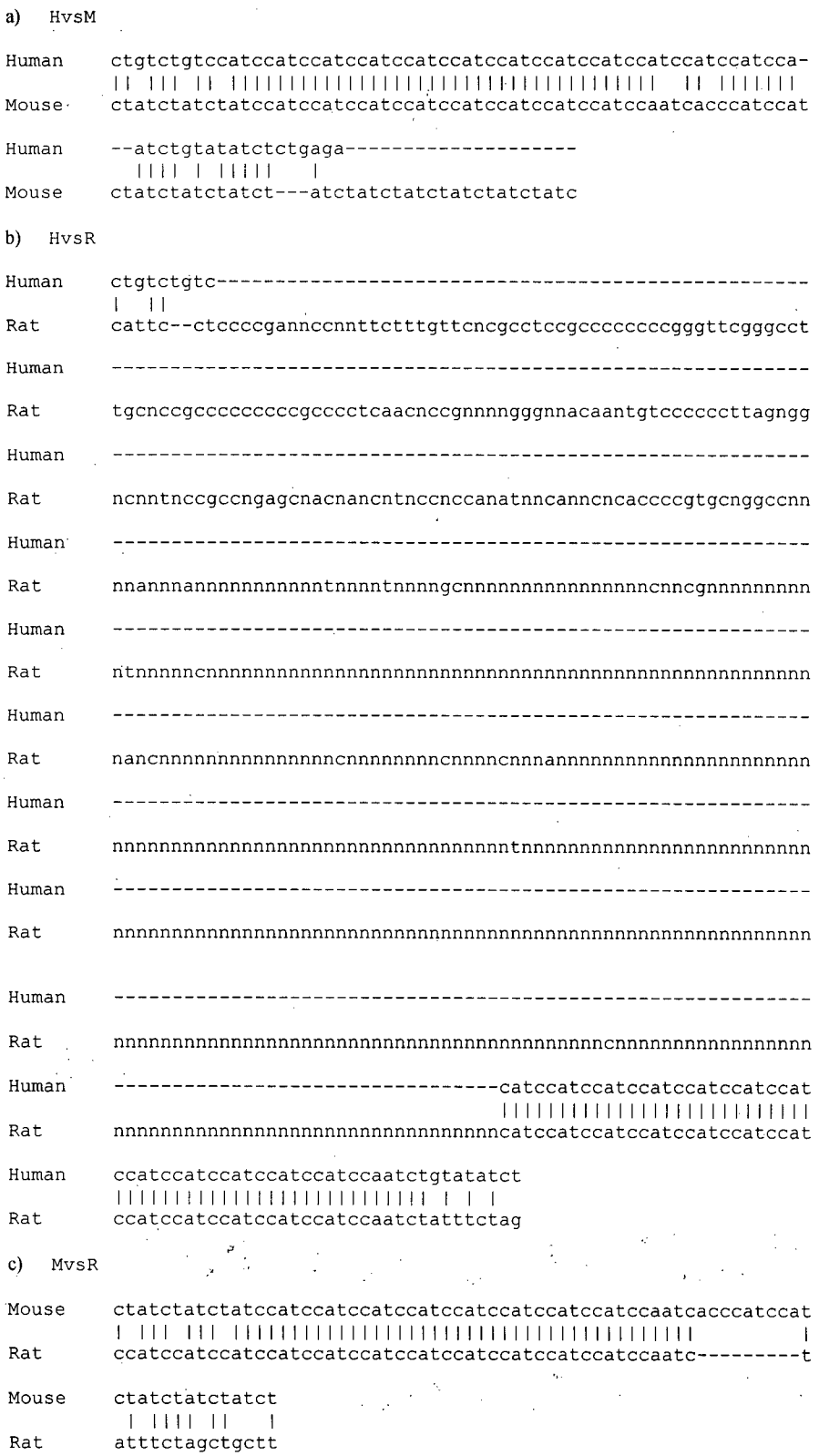


Figure 42. Alignment from mVista around EI1-HHR4 from (a) HvsM, (b) HvsR, and (c) MvsR.

Figure 43. Since this is a repeat region, it is not likely to be functional as a binding site for tfactors, so no further analysis was performed.

3.11 Summary for EI1

The data for EI1 can be summarized as follows (locations within brackets are positions relative to the translation start site (+1)). Four highly homologous regions in EI1 were found. EI1-HHR1 is at Hchr17:29272543-29272633 (+318 to +408), Mchr11:80125572-80125667 (+281 to +376), and Rchr10:61761246-61761340 (+279 to +373). EI1-HHR2 is at Hchr17:29281291-29281430 (+9066 to +9205), Mchr11:80132603-80132753 (+7312 to +7462), and Rchr10:61769227-61769379 (+8260 to +8406). EI1-HHR3 is at Hchr17:29299920-29299983 (+27695 to +27758), Mchr11:80151206-80151279 (+25915 to +25988), and Rchr10:61786728-61786801 (+25761 to +25834). EI1-HHR4 is at Hchr17:29323319-29323389 (+51094 to +51164), Mchr11:80161561-80161634 (+36270 to +36343), and Rchr10:61795122-61795754 (+34155 to +34787).

Within EI1-HHR1, no promising tfactor prediction was found because the predictions from MATCH™ and MatInspector did not agree with each other. Within EI1-HHR2, the most important predictions were Pax-2 from MatInspector at Hchr17:29281355-29281377 (+9130 to +9152), Mccchr11:80132678-80132700 (+7387 to +7409), and Rchr10:61769304-61769326 (+8337 to +8359), and Pax-4 from MATCH™ at Hchr17:29281360-29281380 (+9135 to +9155), Mchr11:80132683-80132703 (+7392 to +7412), and Rchr10:61769309-61769329 (+8342 to +8362). These two predictions are similar and their locations overlap. Within EI1-HHR3, both programs predict a tfactor binding site for MyoD at Hchr17:29299957-29299971 (+27732 to

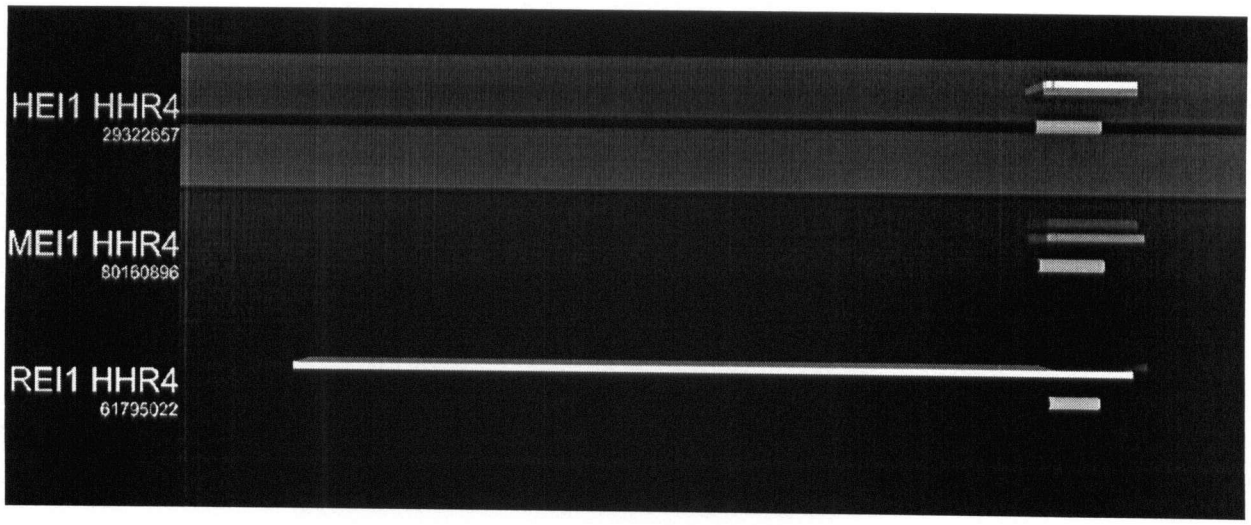


Figure 43. Sockeye presentation of regions surrounding EI1-HHR4 at Hchr17:29322657-29323503, Mchr11:80160896-80161742, and Rchr10:61795022-61795868. Blue rods denote introns. Purple blocks represent repeat sequence. Bars above the introns represent the sequences for EI1-HHR4 with homologies denoted by the following colours: Green – HvsM, Grey – HvsR, Red – MvsR. Note the difference in length for HvsR in the rat sequence because of the gap within the alignment.

+27746), Mchr11:80151244-80151258 (+25953 to +25967), and Rchr10:61786765-61786779 (+25798 to +25812). Because EI1-HHR4 is a simple (TAAC)n repeat at Hchr17:29323326-29323377 (+51101 to +51152), Mchr11:80161568-80161619 (+36277 to +36328), and Rchr10:61795703-61795742 (+34736 to +34775), a functional role as a tfactor binding site is unlikely; and tfactor predictions were not made. Unlike 5UR, no promising homologous regions were detected between *Fugu* and the three mammalian species. Some statistics concerning the length of the highly homologous regions and their tfactor predictions are summarized in Table 26.

Table 26. Summary of the findings from this study. Lengths (in bp) of different regions and highly homologous sequences are presented together with the number of tfactor predictions made by MATCH™ and MatInspector. Individual = initial number of predictions in the segment extending 100 bp upstream to 100 bp downstream of the highly homologous sequence (except in the segment for 5UR-HHR3, which was extended from 200 bp upstream to 200 bp downstream). Shared = numbers of predictions that are shared by human and at least one other species. Total is the total number of nucleotides in the highly homologous sequences or the total number of tfactor predictions (data from NF1HCS and EI1-HHR4 are excluded from the calculation).

	Human					Mouse					Rat				
	Length	MATCH		MatInspector		Length	MATCH		MatInspector		Length	MATCH		MatInspector	
		Individual	Shared	Individual	Shared		Individual	Shared	Individual	Shared		Individual	Shared	Individual	Shared
5UR	59756	NA		NA		59756	NA		NA		59756	NA		NA	
5UR-HHR1	67	76	26	36	7	72	56	25	22	7	68	58	20	22	2
5UR-HHR2	50	28	8	14	2	50	35	7	12	1	50	40	8	16	2
5UR-HHR3	287	79	45	34	31	287	73	43	35	29	286	63	34	37	25
NF1HCS	24	2	2	1	1	24	2	2	1	1	24	2	2	1	1
Total	404	183	79	84	40	409	164	75	69	37	404	161	62	75	29
	Length	MATCH		MatInspector		Length	MATCH		MatInspector		Length	MATCH		MatInspector	
		Individual	Shared	Individual	Shared		Individual	Shared	Individual	Shared		Individual	Shared	Individual	Shared
Intron 1	60613	NA		NA		44115	NA		NA		43352	NA		NA	
EI1-HHR1	91	40	15	32	11	96	48	13	31	11	95	40	15	23	10
EI1-HHR2	140	83	42	39	15	151	94	40	41	12	153	102	37	39	15
EI1-HHR3	64	57	21	16	2	74	61	20	21	2	74	63	20	21	2
EI1-HHR4	71	NA	NA	NA	NA	71	NA	NA	NA	NA	683	NA	NA	NA	NA
Total	295	180	78	87	28	321	203	73	93	25	322	205	72	83	27

Chapter 4. Discussion

This discussion is divided into four sections. First, results related to the *NF1* promoter and additional TSSs are discussed. Then discussion is provided regarding NF1HCS, which is the major discovery of this study. This is followed by some comments on the strengths and limitations of this research. Lastly, some ideas and hopes for future research are presented.

4.1 Defining the *NF1* Promoter Region and Core Promoter Element

In this study, a region of almost 60 kb upstream of the *NF1* translation start site (including 483 bp downstream of the TSS) and all of intron 1 in three mammalian species were analysed for transcriptional control elements using phylogenetic footprinting and other bioinformatic methods. Particular emphasis was placed on efforts to define the *NF1* core promoter element.

The characterization of -484 bp as the major TSS for *NF1* is based on two separate studies (Marchuk *et al.*, 1991; Hajra *et al.*, 1994). Marchuk *et al.* collected cDNA from adult brain and fetal brain, fetal muscle, and endothelial tissue, as well as from a melanoma. These investigators performed primer extension assays beginning 92 bp upstream of the translation start site to identify mRNA products of the *NF1* gene. The longest and major product found had an approximate length of 380 bp to 410 bp, predicting a TSS between 471 bp to 501 bp upstream of the translation start site. Another product was around 300 bp long, which put the TSS at roughly 392 bp upstream of the translation start site. These positions were only approximate; complete sequencing of the 5' cDNA was not done.

Hajra *et al.* (1994) constructed a riboprobe that included 309 bp of *NF1* sequence extending 520 bp to 212 bp upstream of the human *NF1* translation start site. RNase protection assays were then done on cDNA prepared from human brain tissue. A major product of 272 bp was obtained, putting the major TSS at 484 bp (212 + 272 bp) upstream of the translation start site. Other products indicated alternate TSSs at 483 bp and 495 bp upstream of the translation start site. Because the TSS locations established by Hajra *et al.* agreed with the range found in the Marchuk *et al.* studies, the location of major TSS at –484 bp relative to the translation start site is credible.

No TATA box, CAAT box or other core promoter element has been reported for the human *NF1* gene (Viskochil, 1998). The findings in my study confirm and expand this observation -- 5UR-HHR3, which includes the major TSS and extends between 512 bp and 226 bp upstream of the translation start site, has no recognized core promoter element in any of the species analyzed (Tables 11 and 12). Therefore, two questions arise: First, does *NF1* have a core promoter element? Second, if *NF1* does have a core promoter element, where is it?

One way to estimate the likelihood that the human *NF1* gene has a core promoter element is to compare *NF1* with other human genes. A core promoter element is defined as a DNA sequence that interacts directly with the basal transcription machinery (Smale *et al.*, 2003), and only 4 well-defined eukaryotic core promoter elements have been described -- TATA, Inr, BRE, and DPE. The CAAT box and GC box are often associated with promoter regions but are not core promoter elements because they do not interact directly with the basal transcription machinery (Strachan *et al.*, 1999a).

A bioinformatics study of the promoter regions of 1031 human genes showed that only 32% contained TATA boxes and that 85% contained initiator (Inr) elements (Suzuki *et al.*, 2001). This analysis did not include BRE, an upstream element associated with the TATA box, or DPE, a downstream core promoter element that is frequently seen in *Drosophila* genes (Kutach *et al.*, 2000). Suzuki *et al.* (2002) found that the promoter regions of about 15% of human genes had no recognizable TATA or Inr. The percentage of genes with no recognizable core promoter element would probably be smaller if BRE and DPE had also been included in the search. No comparable survey of well-defined core promoter element usage in mouse, rat of *Fugu* has been published.

Because the *NF1* gene is ubiquitously expressed, it has been suggested that the *NF1* promoter is embedded within a CpG island like many other TATA-less housekeeping genes (Viskochil, 1998). According to Goldenpath, a CpG island is found upstream of the *NF1* gene at Hchr17:29271495-29271965, which extends –731 bp to –261 bp relative to the translation start site. Therefore, this CpG island spans the TSS and the promoter region. Both mouse and rat have CpG islands upstream of their translation start sites. Suzuki *et al.* (2002) found that 1/6 of human genes with promoter-associated CpG islands do not have TATA or Inr as a core promoter element. DPE may be found with or without a CpG island, but it is not known how frequently DPE occurs as the only core promoter element in association with a CpG island. There is no Inr associated with the CpG island in the human *NF1* gene (Table 11 and 12). TRANSFAC does not include a DPE consensus sequence, but *NF1* does not have a DPE consensus sequence at the location it would be expected 28 bp to 33 bp downstream of the major TSS.

The promoter regions of many housekeeping genes are associated with a CpG island (Kundu *et al.* 1999), often without a TATA, Inr, or DPE core promoter element (Smale *et al.*, 2003). When there is no recognizable core promoter element within the CpG island, it is not clear how

transcription is initiated or where the basal transcription machinery binds (Smale *et al.*, 2003). CpG islands usually contain multiple Sp1 sites, and it has been hypothesized that Sp1 directs the basal transcription machinery to a particular region within a TATA-less CpG island (Smale, 1994). Within this region, the transcription machinery may choose a window of DNA sequence that is most compatible with its DNA binding motif and begin transcription. The sequence where the basal transcription complex binds in association with a CpG island may be gene specific and may not have a motif that is conserved or recognizable as a core promoter element in other genes. However, this mechanism of CpG-associated transcription initiation has not been demonstrated experimentally, and it is not known whether such facilitative transcription initiation sequences are highly conserved among different species.

Several lines of indirect evidence suggest that additional core promoter elements exist that have not yet been characterized. About 3% of human genes without a CpG island have no TATA or Inr element (Suzuki *et al.*, 2001). Some of these genes may use DPE as a core promoter element, but the mechanism for transcription initiation in the promoter regions of genes that lack both a CpG island and a recognized core promoter element is not clear. Eukaryotic organisms have a set of so-called "TBP-like" proteins (TLP/TRF2/TLF) that are similar to TATA-binding protein (TBP) but are not part of the standard basal transcription machinery (Martinez 2001). TRF2 protein has TATA-binding and transcription activation properties (Hansen *et al.*, 1977). TLP can increase basal transcription from TATA-less promoters (Ohbayashi *et al.*, 2003) but does not bind to the TATA consensus sequence or direct transcription from TATA-containing promoter *in vitro* in mouse (Ohbayashi *et al.*, 1999). Although the mechanism is not clear, TLP may be part of an unconventional transcription initiation process. Recently, a bioinformatic study isolated several novel motifs that are enriched in *Drosophila* promoter regions but are not associated with known core promoter elements (Ohler *et al.*, 2002). These motifs may represent new

transcription factor binding sites or core promoter elements. If additional core promoter elements remain to be discovered, might the *NF1* gene possess such a novel core promoter element?

Most studies on transcription control of the human *NF1* gene have focused on a region upstream of the *NF1* transcription start site (Horan *et al.,* 2000; Mancini *et al.,* 1999; Luijten *et al.,* 2000), but important transcription regulatory elements, including the core promoter element itself, can also occur downstream of a gene's TSS (Kadonaga et al., 2002). The proportion of human genes that contain core promoter elements that are downstream of the TSS is not known, but DPE, a downstream core promoter, has been found in the human *IRF1*, *TAF7*, and *CCR3* genes (Burke *et al.*, 1997; Zhou *et al.*, 2001; Vijh *et al.*, 2002). In *Drosophila,* DPE was found in the promoter region of up to 40% of 205 genes that were analyzed and was the only recognized core promoter element besides Inr in 26% of these genes (Kutach *et al.*, 2000). DPEs are generally associated with Inr with strict spacing between the elements (Kutach *et al.*, 2000).

Two *in vitro* studies of the *NF1* promoter region using the luciferase assay support the possibility that the *NF1* core promoter element lies downstream of the TSS (Figure 15; Purandare *et al.* 1996; Rodenhiser *et al.,* 2002). Purandare *et al.* used a basal luciferase construct that included the portion of the *NF1* 5UR between –4361 bp and -11 bp. (To facilitate comparisons, all nucleotide positions are given relative to the *NF1* translation start site unless otherwise stated). These investigators found that a segment between 341 bp and 11 bp upstream of the *NF1* translation start site can function independently as a promoter but that deletion of this region from a larger construct increased luciferase activity by 65 fold. This observation suggests that a strong repressor is also present in this region. Rodenhiser *et al.* (2002) showed that a construct that includes the segment between –755 bp and –255 bp possesses the highest activity, suggesting the presence of a core promoter and/or a strong enhancer in this region. Shortening

the construct to –330 bp or lengthening it to –131 bp both led to decreased activity. Therefore, a repressor may be located downstream of –255 bp while a core promoter and/or enhancer may be present between –330 bp and –255 bp. Both luciferase assay studies also indicate that the addition of sequence upstream of the major TSS at -484 bp can increase transcriptional activity (Figure 15). These data are consistent with the possibility that NF1HCS is a core promoter element, that NF1HCS is a strong transcriptional activator, or that NF1HCS has both of these functions.

Both studies are consistent with the presence of the *NF1* core promoter between 341 and 255 bp upstream of the translation start site. Note that -341 is chosen over -330 because the segment between -341 bp and -11 bp can function as an independent promoter. This region (-341 to –255 bp) is 154 to 229 bp downstream of the major TSS at –484 bp and overlaps with a large portion of the CpG island at -731 to –261 bp.

PromoterInspector and DPF were used in this study to search for the *NF1* promoter region. PromoterInspector predictions are based on existing annotation and the recognition of a combination of features characteristic of a promoter region (Scherf *et al.*, 2000). PromoterInspector reported a region of 601 bp as the *NF1* promoter (Table 4, Figure 19). This prediction extends from –771 to –111 bp relative to the translation start site.

DPF predicts the TSS for a gene and defines the promoter region as extending 250 bp upstream and 50 bp downstream from the TSS prediction (Bajic *et al.*, 2002). The two TSSs predicted by DPF are at –384 bp and –116 bp, which together predict a potential promoter region that extends from –534 to –85 bp from the ORF. Note that these predictions do not agree with the major TSS established by experiment at –484 bp from the translation start site).

Both programs predict a promoter region for *NF1* that spans the major TSS and includes a segment downstream from this point that contains NF1HCS (Figure 20). These predictions are compatible with the luciferase data and with the possibility that the *NF1* core promoter element lies downstream, rather than upstream, of its major TSS.

4.2 Major Discoveries of This Research

There are three important discoveries from this study.

First, this research has defined three highly homologous regions in the 5UR of *NF1*. 5UR-HHR1 lies in the middle of an intragenic region (Figure 28). Although it is over 42000 bp upstream from the *NF1* gene, it may have an effect on *NF1* transcription because transcription regulatory factors like enhancers are known to act over a long range (Blackwood *et al.*, 1998). For example, the locus control region, which lies roughly 26 kb upstream of the human β-globin gene, acts as an enhancer (Li *et al.*, 1990; Shen *et al.*, 2002), and the enhancer for the human immunoglobulin Calpha1 (IgH-k1 and IgH-k2) genes lies 25 kb downstream (Mills *et al.*, 1997). However, by the same reasoning, 5UR-HH1 could also be an enhancer for a gene that lies further upstream. Portions of 5UR-HH1 are found elsewhere in the human genome, so this sequence may contain functional motifs such as enhancer elements that are also used by other genes. Alternatively, 5UR-HHR1 may be present at this location for another purpose -- for example, as part of a gene that has not yet been detected. There is no gene annotation or GenScan prediction in this region, but UCSC annotates a spliced-out portion of EST AA416617 spanning 5UR-HHR1.

My identification of 5UR-HHR2 (-689 bp to -640) and 5UR-HHR3 (-519 bp to –233 bp) is consistent with what has previously been reported regarding the high homology between the human and mouse genomes in the *NF1* 5' upstream region (Hajra *et al.*, 1994). In addition, my study has shown that similarly high homology (>0.90) exists in this region between human and rat (Figures 25, 30, and 32). Both 5UR-HHR2 and 5UR-HHR3 are likely to be functional because of their high homology among the mammalian species and their proximity to the major *NF1* TSS (at -484 bp).

5UR-HHR3 is probably more important than 5UR-HHR2 for three reasons. First, 5UR-HHR3 is closer to the TSS. Second, 5UR-HHR3 is a longer region with more promising transcription factor binding site predictions. Third, a portion of 5UR-HHR3 (NF1HCS) is conserved among all four species, while no homology between 5UR-HHR2 and the *NF1* 5UR was found in *Fugu*.

The second major finding of this study is the existence of three highly homologous regions in EI1. (A fourth region, EI1-HHR4, was also identified but is comprised of repeat elements and is therefore unlikely to be involved in *NF1* transcriptional regulation.) These homologies have not been reported previously. According to Waterson *et al.* (2002), only 40% of the human genome can be aligned to the mouse genome at the nucleotide level. Furthermore, the average homology of known regulatory regions is only 75.4%, and the average homology for coding regions is 85.7%. Since the identities of all of the highly homologous regions found in intron 1 of *NF1* are more than 90% conserved, they are highly unlikely to be due to chance.

The function of these highly homologous regions is unknown, but they may be related to transcriptional regulation. The transcription factor binding sites predicted in EI-HHR2 and EI-

HHR3 are shared by all three mammalian species (Figures 39 and 41; Tables 22, 23, 24, and 25). Although functional transcription factor binding sites have not previously been described within introns of the *NF1* gene, transcriptional control elements have been found within introns of several other human genes. For example, upregulation of the *Fra-1* gene by an AP-1 site within the first intron has been reported (Casalino *et al.*, 2003). Sequences within introns may also affect splicing. For example, a splicing enhancer has been discovered in intron 6 of the Survival Motor Neuron (*SMN1* and *SMN2*) genes (Miyajima *et al.*, 2002).

Partial matches to all of the highly homologous regions discovered in this study were found in human chromosomes other than chromosome 17 by BLAST searches (data not shown). This is consistent with a potential role of the highly homologous regions in transcriptional control because most transcription factor binding sites occur throughout the genome (Brigg *et al.*, 1986; Whitmarch *et al.*, 1999). Alternatively, the highly homologous regions found in this study could be related to replication or chromatin structure. Identification of these functions would require *in vitro* experiments that are beyond the scope of this study.

The most important discovery of this study is NF1HCS. This 24 bp sequence **acttccggtggggtgtcatggcgg** is located at Hchr17:29271893-29271916 (-333 to -310 bp in relation to the translation start site), Mchr11:80124966-80124989 (-326 to -303 bp), Rchr10:61760813-61760836 (-155 to -132 bp), and FCAAB01003481:22551-22574 (-179 to -156 bp). This sequence is identical in all three mammalian species studied and varies by only a single nucleotide between mammals and *Fugu*. If the mammalian NF1HCS sequence is used in a BLAST query of the human, mouse, or rat genomes, the Expect values found are $7x10^{-5}$ for human, $6x10^{-5}$ for mouse, and $7x10^{-5}$ for rat. If the *Fugu* sequence is used in a BLAST search of the *Fugu* genome, the Expect value is $9x10^{-6}$. This means that the alignments expected for

NF1HCS among these various species are not likely to be due to chance. Because of its high homology and location near the TSS, NF1HCS is probably related to *NF1* transcriptional control.

I was unable to determine the function of NF1HCS by locating the sequence in relationship to other vertebrate or *Drosophila* genes using BLAST (Table 13). Comparison with the Rfam and SCOR databases did not indicate any potential secondary or tertiary structure of importance within the NF1HCS RNA. Pairwise BLAST between mammalian NF1HCS and the EPD or TRRD database provided no significant alignment. However, promoter regions are represented in EDP by a segment that only extends 100 bp downstream of the major TSS while NF1HCS lies 150 bp downstream of the *NF1* TSS. The TRRD database is relatively small and does not yet include data from *Fugu*.

Portions of NF1HCS were found occasionally in various regions of the genomes of various organisms, but these locations were not consistent among species. However, exact 24 bp matchs would not necessarily be expected to occur in other locations even if NF1HCS contains a transcription factor binding site or core promoter element. Transcription factor binding sites generally have rather variable consensus sequences (Roulet *et al.*, 1998) that can be very short. For example, the consensus sequence for the TATA box and the consensus sequence for Sp1 site are both only 6 bp long (Kadonaga, 2002; Cook *et al.*, 1999). MATCH^{TM} and MatInspector, the two tfactor binding site detection programs used in this study, define their matrices for transcription factor binding site core sequences with 5 bp and 4 bp, respectively. Therefore, even if only a portion of NF1HCS were found elsewhere in vertebrate genomes, these portions might be long enough to carry out transcriptional regulatory functions.

Some transcription factors like AP-1 can function downstream, upstream or within an intron of a gene (Aringer *et al.*, 2003; Casalino *et al.*, 2003). Most of the BLAST hits for portions of the NF1HCS sequence in other vertebrate genomes lie within an intron of an annotated gene or gene prediction. This location in the non-coding segment of a gene is compatible with a *cis*-regulatory function for these sequences (Ureta-Vidal *et al.*, 2003). Although no promising predictions shared by MATCH^{TM} and MatInspector for known transcription factor binding sites were found within NF1HCS, this sequence may contain one or more novel tfactor binding sites.

Phylogenetic footprinting of the promoter regions of 5 other genes with known core promoter elements in human, mouse, rat and *Fugu* was undertaken to help assess the biological significance of the very high homology found for NF1HCS among these four species. Analysis of the regions surrounding the core promoter elements of the *HBB, ACTA1, TFAP2C, TAF7,* and *LCK* genes revealed various levels of homology. Some of these examples showed a degree of identity as great as that found in NF1HCS among human, mouse, and rat but not with *Fugu.*

Overall, these analyses are far from definitive, but they do provide some support for the notion that NF1HCS functions in the regulation of *NF1* transcription.

As mentioned in the beginning of the Discussion, my analysis indicates that the most likely location for an *NF1* core promoter element is between –255 bp and –341 bp. This region is downstream of the major TSS at –484 bp, overlaps the CpG island, and includes NF1HCS, which lies between -333 bp and -310 bp. The presence of a core promoter in this region is consistent with the predictions of both PromoterInspector and DPF (Figure 21). However, if NF1HCS is the *NF1* core promoter element, it lies further downstream of the TSS than any other core promoter that has ever been described.

The strongest support for the potential function of NF1HCS as the *NF1* core promoter element comes from the luciferase assay experiments (Figure 15; Purandare *et al.* 1996; Rodenhiser *et al.*, 2002). The luciferase assay data are compatible with a core promoter element in the region where NF1HCS lies but are not compatible with a core promoter element in the conventional location between 100 bp upstream and 50 bp downstream of the TSS (Figure 15). Whether NF1HCS or some other portion of the *NF1* promoter region interacts directly with the basal transcription machinery (and is, by definition, a core promoter element) can only be determined through future *in vitro* experiments.

In summary, if the *NF1* gene has a core promoter element, it may be NF1HCS. No TATA box or other recognized core promoter element is present in the vicinity of the *NF1* TSS, and no alternate candidate for the core promoter element was found within 100 bp downstream or 50 bp upstream of major TSS, where all recognized vertebrate core promoter elements lie (Smale *et al.*, 2003).

4.3 Limitations and Strengths of This Research

This study relies heavily on mVista for sequence alignment. As demonstrated by EI1-HHR4, mVista can produce alignments as a result of repeat sequences. Furthermore, finding mVista alignments between *Fugu* and mammalian species is very difficult because of the evolutionary distance and sequence length differences between these species. These problems cannot be compensated by Frameslider, which is dependent on mVista, or by Pairwise BLAST, which often produces numerous hits under low stringency. Although NF1HCS was successfully

located after combining all of these alignment methods, other important homologous regions may exist that were not detected.

The presence of gaps can interfere with mVista alignments of homologous regions. For example, EI1-HHR3 was not initially detected in human and rat because of the presence of a large gap in the alignment. Multiple methods of alignment were done in this study, and these different approaches tend to complement each other. It is, therefore, likely that the highly homologous regions identified are valid.

The cutoff values used for Frameslider in this study were very high. Transcription factor binding sites, unlike amino acid sequences in proteins or restriction enzyme recognition sites in DNA, are very tolerant of variation in sequence (Roulet *et al.*, 1998). Therefore, regions of lower homology may contain many important transcription factor binding sites that would not have been identified as highly homologous regions in this study.

MATCH, MatInspector, and all other currently-available programmes for identifying transcription factor binding sites are inefficient, with many false-positive predictions (Roulet *et al.*, 1998). Even by restricting the analysis to highly homologous regions and considering only predictions that are seen at the same site in more than one species, many of the of predicted tfactor binding sites are probably incorrect. To make matters worse, predictions from MATCH^{TM} and MatInspector usually differ from each other. In this study, there were only two identical predictions among all 4 species studied (AP-1 in 5UR-HHR1 and MyoD in EI1-HHR3) out of 119 shared predictions based on alignment. Experimental validation of predicted tfactor binding site is always necessary. Furthermore, these programs can only predict known transcription factor binding sites. Current knowledge of transcription regulation in mammals is still very

limited, and many transcription factors and regulatory pathways probably have yet to be discovered. It is quite possible that some of the highly-homologous regions identified in this study have functions that are unrelated to transcriptional regulation.

I analysed only the 5UR and intron 1 of the human *NF1* gene in this study, and important transcriptional regulatory factors may exist further upstream, in other introns, in the 3' UTR or further downstream of the gene. There may also be other kinds of regulation (*e.g.*, regulation of chromatin structure, methylation, *etc.*) that were not studied at all.

Despite these limitations, this study has shown that phylogenetic footprinting is a powerful means of discovering regions that are potentially important in transcriptional regulation. As shown in Table 26, 699 out of 120369 bp in the non-coding segments of the human *NF1* 5UR and intron 1 sequence were found to have homologies with mouse and rat that extend for at least 50 bp and are as strong as those that occur in the coding regions of this gene. In some of these cases, the homologies in the non-coding sequence are as strong as those observed with the *NF1* coding regions of *Fugu* as well. If most of the sites that are critical to *NF1* transcriptional regulation lie in these highly homologous regions, application of phylogenetic footprinting will have reduced the search space for important transcriptional control regions by more than 170-fold. Within these regions, 225 out of the 534 transcription factor predictions made in human were also found in mouse and/or rat, thereby decreasing the number of predictions that are most likely to be valid by an additional 60%. Phylogenetic footprinting has been used for a variety of genes in various organisms (Hong *et al.*, 2003, Cliften *et al.*, 2003), and programs have been developed to facilitate this approach (Lenhard *et al.*, 2003). Focusing experimental studies on regions that are shown by phylogenetic footprinting and tfactor binding site analysis to be the

best candidates is likely to be a very efficient strategy for defining the transcriptional regulation of *NF1* (Duret *et al.*, 1997).

This study has increased our understanding of transcriptional regulation of the human *NF1* gene in two ways. First, if the tfactor binding sites predicted within the highly homologous regions of the 5UR and intron 1 are true, their further study may elucidate important aspects of transcriptional control of the *NF1* gene. Screening on these locations may identify disease-causing mutations in NF1 patients in whom no mutation of the coding sequence or splicing has been found (Upadhyaya *et al.*, 1994; Fahsold *et al.*, 2000; Messiaen *et al.*, 2000). Furthermore, since many of the clinical manifestations of NF1 may be related to haploinsufficiency (McLaughlin *et al.*, 2002), up-regulation of *NF1* gene expression by activating specific transcriptional regulatory pathways may provide a novel approach to treatment of the disease.

Second, if NF1HCS is the *NF1* core promoter element, it is also an important target for mutational screening, since mutations in the promoter region may disrupt transcription initiation (Duan *et al.*, 2002). Furthermore, if NF1HCS contains a core promoter or other important transcriptional regulatory element, it may also occur in other genes.

This study has shown that phylogenetic footprinting can be used to compare non-coding regions for the *NF1* gene in different organisms. The same method should be applied to other introns and the 3' UTR of *NF1*, which may contain other important regulatory motifs. These methods are likely to be useful in studying the transcriptional regulation of other genes as well.

4.4 Future Ideas and Hopes

The completion of sequencing of many different eukaryotic genomes signals the beginning of an era of comparative study. By using comparisons to more organisms, more homologous regions can be identified with more confidence. For example, in this study, the evolutionary distance between the mammalian species and *Fugu* may be too large for some comparisons. If the sequences of intermediate vertebrates that are more similar to the mammalian species were available, phylogenetic footprinting might be more informative (Thomas *et al.,* 2002). Similarly, information from a species that is more closely related to humans than mice and rats may also be useful. Neurofibromatosis 1 does not occur naturally as a disease in mice or rats, and, although transgenic models are very useful, these animals display a different spectrum of phenotypes in association with constitutional *NF1* mutations than humans with NF1. The extent to which these differences reflect differences in transcriptional regulation of the *NF1* gene is unknown. Transcriptional regulation of *NF1* is more likely to be similar in humans and non-human primates than in humans and rodents.

Only limited experimental analysis of transcriptional regulation of the *NF1* gene has been reported (Hajra *et al.,*1994; Purandare *et al.* 1996; Rodenhiser *et al.,* 2002), and additional experimental studies *in vitro* or in other model systems are necessary. Detailed *in vitro* analysis of NF1HCS is especially important because this region may be crucial for *NF1* transcription initiation and/or regulation. Analysis of this region in knockout mice may also be informative.

Analysis of the highly homologous regions, and especially of NF1HCR, for constitutional mutations in human NF1 patients in whom no coding sequence or splice site mutation can be found may provide evidence regarding the functional importance of the regions identified in this

study. Given that NF1 is such a prevalent genetic disease and that its impact on patients' lives is so great, more research on NF1 is much needed.

Bibliography

Ainsworth P, Rodenhiser D, Stuart A, Jung J. Characterization of an intron 31 splice junction mutation in the neurofibromatosis type 1 (NF1) gene. Hum Mol Genet. 1994 Jul;3(7):1179-81.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10.

Annilo T, Chen ZQ, Shulenin S, Dean M. Evolutionary analysis of a cluster of ATP-binding cassette (ABC) genes. Mamm Genome. 2003 Jan;14(1):7-20.

Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, Christoffels A, Rash S, Hoon S, Smit A, Gelpke MD, Roach J, Oh T, Ho IY, Wong M, Detter C, Verhoef F, Predki P, Tay A, Lucas S, Richardson P, Smith SF, Clark MS, Edwards YJ, Doggett N, Zharkikh A, Tavtigian SV, Pruss D, Barnstead M, Evans C, Baden H, Powell J, Glusman G, Rowen L, Hood L, Tan YH, Elgar G, Hawkins T, Venkatesh B, Rokhsar D, Brenner S. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science. 2002 Aug 23;297(5585):1301-10.

Aringer M, Hofmann SR, Frucht DM, Chen M, Centola M, Morinobu A, Visconti R, Kastner DL, Smolen JS, O'Shea JJ. Characterization and Analysis of the Proximal Janus Kinase 3 Promoter. J Immunol. 2003 Jun 15;170(12):6057-6064.

Ars E, Serra E, Garcia J, Kruyer H, Gaona A, Lazaro C, Estivill X. Mutations affecting mRNA splicing are the most common molecular defects in patients with neurofibromatosis type 1. Hum Mol Genet. 2000 Jan 22;9(2):237-47.

Attwood JT, Yung RL, Richardson BC. DNA methylation and the regulation of gene transcription. Cell Mol Life Sci. 2002 Feb;59(2):241-57.

Bajenaru ML, Donahoe J, Corral T, Reilly KM, Brophy S, Pellicer A, Gutmann DH. Neurofibromatosis 1 (NF1) heterozygosity results in a cell-autonomous growth advantage for astrocytes. Glia. 2001 Mar 15;33(4):314-23.

Bajic VB, Seah SH, Chong A, Zhang G, Koh JL, Brusic V. Dragon Promoter Finder: recognition of vertebrate RNA polymerase II promoters. Bioinformatics. 2002 Jan;18(1):198-9.

Basu TN, Gutmann DH, Fletcher JA, Glover TW, Collins FS, Downward J. Aberrant regulation of ras proteins in malignant tumour cells from type 1 neurofibromatosis patients. Nature. 1992 Apr 23;356(6371):713-5.

Beato M, Eisfeld K. Transcription factor access to chromatin. Nucleic Acids Res. 1997 Sep 15;25(18):3559-63

Benish BM. Letter: "The neurocristopathies: a unifying concept of disease arising in neural crest development". Hum Pathol. 1975 Jan;6(1):128.

Bhawan J, Purtilo DT, Riordan JA, Saxena VK, Edelstein L. Giant and "granular melanosomes" in Leopard syndrome: an ultrastructural study. J Cutan Pathol. 1976;3(5):207-16.

Black AR, Black JD, Azizkhan-Clifford J. Sp1 and kruppel-like factor family of transcription factors in cell growth regulation and cancer. J Cell Physiol. 2001 Aug;188(2):143-60.

Blackwood EM, Kadonaga JT. Going the distance: a current view of enhancer action. Science. 1998 Jul 3;281(5373):61-3.

Brenner S, Venkatesh B, Yap WH, Chou CF, Tay A, Ponniah S, Wang Y, Tan YH. Conserved regulation of the lymphocyte-specific expression of lck in the Fugu and mammals. Proc Natl Acad Sci U S A. 2002 Mar 5;99(5):2936-41. Epub 2002 Feb 26.

Briggs MR, Kadonaga JT, Bell SP, Tjian R. Purification and biochemical characterization of the promoter-specific transcription factor, Sp1. Science. 1986 Oct 3;234(4772):47-52.

Burke TW, Kadonaga JT. Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters. Genes Dev. 1996 Mar 15;10(6):711-24.

Burke TW, Kadonaga JT. Abstract The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. Genes Dev. 1997 Nov 15;11(22):3020-31.

Casalino L, De Cesare D, Verde P. Accumulation of Fra-1 in ras-Transformed Cells Depends on Both Transcriptional Autoregulation and MEK-Dependent Posttranslational Stabilization. Mol Cell Biol. 2003 Jun;23(12):4401-15.

Cawthon RM, Weiss R, Xu GF, Viskochil D, Culver M, Stevens J, Robertson M, Dunn D, Gesteland R, O'Connell P, et al. A major segment of the neurofibromatosis type 1 gene: cDNA sequence, genomic structure, and point mutations. Cell. 1990 Jul 13;62(1):193-201.

Clark SJ, Harrison J, Molloy PL. Sp1 binding is inhibited by (m)Cp(m)CpG methylation. Gene. 1997 Aug 11;195(1):67-71.

Cliften P, Sudarsanam P, Desikan A, Fulton L, Fulton B, Majors J, Waterston R, Cohen BA, Johnston M. Finding Functional Features in Saccharomyces Genomes by Phylogenetic Footprinting. Science. 2003 May 29 [Epub ahead of print]

Comb M, Goodman HM. CpG methylation inhibits proenkephalin gene expression and binding of the transcription factor AP-2. Nucleic Acids Res. 1990 Jul 11;18(13):3975-82.

Cook T, Gebelein B, Urrutia R. Sp1 and its likes: biochemical and functional predictions for a growing family of zinc finger transcription factors. Ann N Y Acad Sci. 1999 Jun 30;880:94-102.

Corden J, Wasylyk B, Buchwalder A, Sassone-Corsi P, Kedinger C, Chambon P. Promoter sequences of eukaryotic protein-coding genes. Science. 1980 Sep 19;209(4463):1406-14.

Cramer P, Srebrow A, Kadener S, Werbajh S, de la Mata M, Meen G, Nogues G, Kornblihtt AR. Coordination between transcription and pre-mRNA processing. FEBS Lett. 2001 Jun 8;498(2-3):179-82.

Crowe FW, Schull WJ, Neel JV. A Clinical, Pathological and Genetic Study of Multiple Neurofibromatosis, Springfield, IL: Charles C. Thomas; 1956

Crowe FW. Axillary freckling as a diagnostic aid in neurofibromatosis. Ann Intern Med 1964;61-1142

Dasgupta B, Gutmann DH. Neurofibromatosis 1: closing the GAP between mice and men. Curr Opin Genet Dev. 2003 Feb;13(1):20-7.

DeBella K, Szudek J, Friedman JM. Use of the national institutes of health criteria for diagnosis of neurofibromatosis 1 in children. Pediatrics. 2000 Mar;105(3 Pt 1):608-14.

DeClue JE, Papageorge AG, Fletcher JA, Diehl SR, Ratner N, Vass WC, Lowy DR. Abnormal regulation of mammalian p21ras contributes to malignant tumor growth in von Recklinghausen (type 1) neurofibromatosis. Cell. 1992 Apr 17;69(2):265-73.

Duan ZJ, Fang X, Rohde A, Han H, Stamatoyannopoulos G, Li Q. Developmental specificity of recruitment of TBP to the TATA box of the human gamma-globin gene. Proc Natl Acad Sci U S A. 2002 Apr 16;99(8):5509-14.

Dugoff, L.; Sujansky, E. Neurofibromatosis type 1 and pregnancy. Am. J. Med. Genet. 66: 7-10, 1996.

Duret L, Bucher P. Searching for regulatory elements in human noncoding sequences. Curr Opin Struct Biol. 1997 Jun;7(3):399-406.

Easton DF, Ponder MA, Huson SM, Ponder BA. An analysis of variation in expression of neurofibromatosis (NF) type 1 (NF1): evidence for modifying genes. Am J Hum Genet. 1993 Aug;53(2):305-13.

Evans DG, Baser ME, McGaughran J, Sharif S, Howard E, Moran A. Malignant peripheral nerve sheath tumours in neurofibromatosis 1. J Med Genet. 2002 May;39(5):311-4.

Evans R, Fairley JA, Roberts SG. Activator-mediated disruption of sequence-specific DNA contacts by the general transcription factor TFIIB. Genes Dev. 2001 Nov 15;15(22):2945-9.

Fahsold R, Hoffmeyer S, Mischung C, Gille C, Ehlers C, Kucukceylan N, Abdel-Nour M, Gewies A, Peters H, Kaufmann D, Buske A, Tinschert S, Nurnberg P. Minor lesion mutational spectrum of the entire NF1 gene does not explain its high mutability but points to a functional domain upstream of the GAP-related domain. Am J Hum Genet. 2000 Mar;66(3):790-818.

Ferner RE, Gutmann DH. International consensus statement on malignant peripheral nerve sheath tumors in neurofibromatosis. Cancer Res. 2002 Mar 1;62(5):1573-7.

Ferner RE, Lucas JD, O'Doherty MJ, Hughes RA, Smith MA, Cronin BF, Bingham J. Evaluation of (18)fluorodeoxyglucose positron emission tomography ((18)FDG PET) in the detection of malignant peripheral nerve sheath tumours arising from within plexiform neurofibromas in neurofibromatosis 1. J Neurol Neurosurg Psychiatry. 2000 Mar;68(3):353-7.

Ferner RE, O'Doherty MJ. Neurofibroma and schwannoma. Curr Opin Neurol. 2002 Dec;15(6):679-84.

Frech K, Quandt K, Werner T. Finding protein-binding sites in DNA sequences: the next generation. Trends Biochem Sci. 1997 Mar;22(3):103-4.

Friedman JM, Birch PH. Type 1 neurofibromatosis: a descriptive analysis of the disorder in 1,728 patients. Am J Med Genet. 1997 May 16;70(2):138-43.

Friedman JM, Riccardi VM. Clinical and Epidermiologic Features. In: Friedman JM, Gutmann DH, MacCollin M, Riccardi VM editors. Neurofibromatosis: Phenotype, Natural History, and Pathogenesis. Baltimore and London: The Johns Hopkins University Press. 1999. pp. 29-87

Friedman JM. Clinical Genetics. In: Friedman JM, Gutmann DH, MacCollin M, Riccardi VM editors. Neurofibromatosis: Phenotype, Natural History, and Pathogenesis. Baltimore and London: The Johns Hopkins University Press. 1999. pp. 110-118

Friedman JM. Neurofibromatosis 1: clinical manifestations and diagnostic criteria. J Child Neurol. 2002 Aug;17(8):548-54.

Gasch AP, Eisen MB. Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering. Genome Biol. 2002 Oct 10;3(11):RESEARCH0059.

Geyer PK, Clark I. Protecting against promiscuity: the regulatory role of insulators. Cell Mol Life Sci. 2002 Dec;59(12):2112-27.

Gillemans N, McMorrow T, Tewari R, Wai AW, Burgtorf C, Drabek D, Ventress N, Langeveld A, Higgs D, Tan-Un K, Grosveld F, Philipsen S. Functional and comparative analysis of globin loci in pufferfish and humans. Blood. 2003 Apr 1;101(7):2842-9. Epub 2002 Nov 27.

Ginty DD, Bonni A, Greenberg ME. Nerve growth factor activates a Ras-dependent protein kinase that stimulates c-fos transcription via phosphorylation of CREB. Cell. 1994 Jun 3;77(5):713-25.

Goode DK, Snell PK, Elgar GK. Comparative analysis of vertebrate Shh genes identifies novel conserved non-coding sequence. Mamm Genome. 2003 Mar;14(3):192-201.

Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 2003 Jan 1;31(1):439-41

Gutmann DH, Aylsworth A, Carey JC, Korf B, Marks J, Pyeritz RE, Rubenstein A, Viskochil D. The diagnostic evaluation and multidisciplinary management of neurofibromatosis 1 and neurofibromatosis 2. JAMA. 1997 Jul 2;278(1):51-7.

Gutmann DH, Cole JL, Collins FS. Expression of the neurofibromatosis type 1 (NF1) gene during mouse embryonic development. Prog Brain Res. 1995;105:327-35.
Gutmann DH, Donahoe J, Brown T, James CD, Perry A. Loss of neurofibromatosis 1 (NF1) gene expression in NF1-associated pilocytic astrocytomas. Neuropathol Appl Neurobiol. 2000 Aug;26(4):361-7.
Gutmann DH, Loehr A, Zhang Y, Kim J, Henkemeyer M, Cashen A. Haploinsufficiency for the neurofibromatosis 1 (NF1) tumor suppressor results in increased astrocyte proliferation. Oncogene. 1999 Aug 5;18(31):4450-9.
Gutmann DH, Wu YL, Hedrick NM, Zhu Y, Guha A, Parada LF. Heterozygosity for the neurofibromatosis 1 (NF1) tumor suppressor results in abnormalities in cell attachment, spreading and motility in astrocytes. Hum Mol Genet. 2001 Dec 15;10(26):3009-16.
Gutmann, DH.; Wood, DL.; Collins, FS. Identification of the neurofibromatosis type 1 gene product. Proc. Nat. Acad. Sci. 88: 9658-9662, 1991.
Haines TR, Rodenhiser DI, Ainsworth PJ. Allele-specific non-CpG methylation of the Nf1 gene during early mouse development. Dev Biol. 2001 Dec 15;240(2):585-98.
Hajra A, Martin-Gallardo A, Tarle SA, Freedman M, Wilson-Gunn S, Bernards A, Collins FS. DNA sequences in the promoter region of the NF1 gene are highly conserved between human and mouse. Genomics. 1994 Jun;21(3):649-52.
Hansen SK, Takada S, Jacobson RH, Lis JT, Tjian R. Transcription properties of a cell type-specific TATA-binding protein, TRF. Cell. 1997 Oct 3;91(1):71-83.
Harr R, Haggstrom M, Gustafsson P. Search algorithm for pattern match analysis of nucleic acid sequences. Nucleic Acids Res. 1983 May 11;11(9):2943-57.
Hasleton MD, Ibbitt JC, Hurst HC. Characterisation of the human AP-2gamma gene: control of expression by Sp1/Sp3 in breast tumour cells. Biochem J. 2003 May 6
Hatta N, Horiuchi T, Watanabe I, Kobayashi Y, Shirakata Y, Ohtsuka H, Minami T, Ueda K, Kokoroishi T, Fujita S. NF1 gene mutations in Japanese with neurofibromatosis 1 (NF1). Biochem Biophys Res Commun. 1995 Jul 17;212(2):697-704.
Hong RL, Hamaguchi L, Busch MA, Weigel D. Regulatory Elements of the Floral Homeotic Gene AGAMOUS Identified by Phylogenetic Footprinting and Shadowing. Plant Cell. 2003 Jun;15(6):1296-1309.
Horan MP, Cooper DN, Upadhyaya M. Hypermethylation of the neurofibromatosis type 1 (NF1) gene promoter is not a common event in the inactivation of the NF1 gene in NF1-specific tumours. Hum Genet. 2000 Jul;107(1):33-9.
Iguchi-Ariga SM, Schaffner W. CpG methylation of the cAMP-responsive enhancer/promoter sequence TGACGTCA abolishes specific factor binding as well as transcriptional activation. Genes Dev. 1989 May;3(5):612-9.
Jacob GA, Kitzmiller JA, Luse DS. RNA polymerase II promoter strength in vitro may be reduced by defects at initiation or promoter clearance. J Biol Chem. 1994 Feb 4;269(5):3655-63
Jadayel D, Fain P, Upadhyaya M, Ponder MA, Huson SM, Carey J, Fryer A, Mathew CG, Barker DF, Ponder BA. Paternal origin of new mutations in von Recklinghausen neurofibromatosis. Nature. 1990 Feb 8;343(6258):558-9.
Jost JP. Nuclear extracts of chicken embryos promote an active demethylation of DNA by excision repair of 5-methyldeoxycytidine. Proc Natl Acad Sci U S A. 1993 May 15;90(10):4684-8.
Kadonaga JT, Carner KR, Masiarz FR, Tjian R. Isolation of cDNA encoding transcription factor Sp1 and functional analysis of the DNA binding domain. Cell. 1987 Dec 24;51(6):1079-90.
Kadonaga JT. The DPE, a core promoter element for transcription by RNA polymerase II. Exp Mol Med. 2002 Sep 30;34(4):259-64.

Kanehisa M, Bork P. Bioinformatics in the post-sequence era. Nat Genet. 2003 Mar;33 Suppl:305-10.
Kehrer-Sawatzki H, Moschgath E, Maier C, Legius E, Elgar G, Krone W. Characterization of the Fugu rubripes NLK and FN5 genes flanking the NF1 (Neurofibromatosis type 1) gene in the 5' direction and mapping of the human counterparts. Gene. 2000 Jun 13;251(1):63-71.
Kent, WJ. BLAT: The BLAST-like Alignment Tool. Genome Research 2002 Apr; 12(4):656-664.
Klosterman PS, Tamura M, Holbrook SR, Brenner SE. SCOR: a Structural Classification of RNA database. Nucleic Acids Res. 2002 Jan 1;30(1):392-4.
Kolchanov NA, Ignatieva EV, Ananko EA, Podkolodnaya OA, Stepanenko IL, Merkulova TI, Pozdnyakov MA, Podkolodny NL, Naumochkin AN, Romashchenko AG. Transcription Regulatory Regions Database (TRRD): its status in 2002. Nucleic Acids Res. 2002 Jan 1;30(1):312-7.
Konrad K, Wolff K, Honigsmann H. The giant melanosome: a model of deranged melanosome-morphogenesis. J Ultrastruct Res. 1974 Jul;48(1):102-23.
Korf BR. Clinical features and pathobiology of neurofibromatosis 1. J Child Neurol. 2002 Aug;17(8):573-7; discussion 602-4, 646-51.
Kourea HP, Cordon-Cardo C, Dudas M, Leung D, Woodruff JM. Expression of p27(kip) and other cell cycle regulators in malignant peripheral nerve sheath tumors and neurofibromas: the emerging role of p27(kip) in malignant transformation of neurofibromas. Am J Pathol. 1999 Dec;155(6):1885-91.
Kourea HP, Orlow I, Scheithauer BW, Cordon-Cardo C, Woodruff JM. Deletions of the INK4A gene occur in malignant peripheral nerve sheath tumors but not in neurofibromas. Am J Pathol. 1999 Dec;155(6):1855-60.
Kumar S, Hedges SB. A molecular timescale for vertebrate evolution. Nature. 1998 Apr 30;392(6679):917-20.
Kutach AK, Kadonaga JT. The downstream promoter element DPE appears to be as widely used as the TATA box in Drosophila core promoters. Mol Cell Biol. 2000 Jul;20(13):4754-64.
Lagrange T, Kapanidis AN, Tang H, Reinberg D, Ebright RH. New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB. Genes Dev. 1998 Jan 1;12(1):34-44.
Lakkis MM, Golden JA, O'Shea KS, Epstein JA. Neurofibromin deficiency in mice causes exencephaly and is a modifier for Splotch neural tube defects. Dev Biol. 1999 Aug 1;212(1):80-92.
Lakkis MM, Tennekoon GI. Neurofibromatosis type 1. I. General overview. J Neurosci Res. 2000 Dec 15;62(6):755-63.
Lau N, Feldkamp MM, Roncari L, Loehr AH, Shannon P, Gutmann DH, Guha A. Loss of neurofibromin is associated with activation of RAS/MAPK and PI3-K/AKT signaling in a neurofibromatosis 1 astrocytoma. J Neuropathol Exp Neurol. 2000 Sep;59(9):759-67.
Lazaro C, Ravella A, Gaona A, Volpini V, Estivill X. Neurofibromatosis type 1 due to germ-line mosaicism in a clinically normal father. N Engl J Med. 1994 Nov 24;331(21):1403-7.
Ledbetter DH, Rich DC, O'Connell P, Leppert M, Carey JC. Precise localization of NF1 to 17q11.2 by balanced translocation. Am J Hum Genet. 1989 Jan;44(1):20-4.
Legius E, Marchuk DA, Collins FS, Glover TW. Somatic deletion of the neurofibromatosis type 1 gene in a neurofibrosarcoma supports a tumour suppressor gene hypothesis. Nat Genet. 1993 Feb;3(2):122-6.
Lenhard B, Sandelin A, Mendoza L, Engstrom P, Jareborg N, Wasserman WW. Identification of conserved regulatory elements by comparative genome analysis. J Biol. 2003 [Epub ahead of print]

Leppig KA, Kaplan P, Viskochil D, Weaver M, Ortenberg J, Stephens K. Familial neurofibromatosis 1 microdeletions: cosegregation with distinct facial phenotype and early onset of cutaneous neurofibromata. Am J Med Genet. 1997 Dec 12;73(2):197-204.

Lewis BA, Kim TK, Orkin SH. A downstream element in the human beta-globin promoter: evidence of extended sequence-specific transcription factor IID contacts. Proc Natl Acad Sci U S A. 2000 Jun 20;97(13):7172-7.

Li QL, Zhou B, Powers P, Enver T, Stamatoyannopoulos G. Beta-globin locus activation regions: conservation of organization, structure, and function. Proc Natl Acad Sci U S A. 1990 Nov;87(21):8207-11

Li Y, O'Connell P, Breidenbach HH, Cawthon R, Stevens J, Xu G, Neil S, Robertson M, White R, Viskochil D. Genomic organization of the neurofibromatosis 1 gene (NF1). Genomics. 1995 Jan 1;25(1):9-18.

Lin SY, Black AR, Kostic D, Pajovic S, Hoover CN, Azizkhan JC. Cell cycle-regulated association of E2F1 and Sp1 is related to their functional interaction. Mol Cell Biol. 1996 Apr;16(4):1668-75.

Listernick R, Darling C, Greenwald M, Strauss L, Charrow J. Optic pathway tumors in children: the effect of neurofibromatosis type 1 on clinical manifestations and natural history. J Pediatr. 1995 Nov;127(5):718-22.

Littler M, Morton NE. Segregation analysis of peripheral neurofibromatosis (NF1). J Med Genet. 1990 May;27(5):307-10.

Luijten M, Redeker S, van Noesel MM, Troost D, Westerveld A, Hulsebos TJ. Microsatellite instability and promoter methylation as possible causes of NF1 gene inactivation in neurofibromas. Eur J Hum Genet. 2000 Dec;8(12):939-45.

Lynch TM, Gutmann DH. Neurofibromatosis 1. Neurol Clin. 2002 Aug;20(3):841-65.

Mancini DN, Singh SM, Archer TK, Rodenhiser DI. Site-specific DNA methylation in the neurofibromatosis (NF1) promoter interferes with binding of CREB and SP1 transcription factors. Oncogene. 1999 Jul 15;18(28):4108-19.

Marchuk DA, Saulino AM, Tavakkol R, Swaroop M, Wallace MR, Andersen LB, Mitchell AL, Gutmann DH, Boguski M, Collins FS. cDNA cloning of the type 1 neurofibromatosis gene: complete sequence of the NF1 gene product. Genomics. 1991 Dec;11(4):931-40.

Martinez E, Chiang CM, Ge H, Roeder RG. TATA-binding protein-associated factor(s) in TFIID function through the initiator to direct basal transcription from a TATA-less class II promoter. EMBO J. 1994 Jul 1;13(13):3115-26.

Martinez E. Multi-protein complexes in eukaryotic gene transcription. Plant Mol Biol. 2002 Dec;50(6):925-47.

Martuza RL, Philippe I, Fitzpatrick TB, Zwaan J, Seki Y, Lederman J. Melanin macroglobules as a cellular marker of neurofibromatosis: a quantitative study. J Invest Dermatol. 1985 Oct;85(4):347-50.

Mathis DJ, Chambon P. The SV40 early region TATA box is required for accurate in vitro initiation of transcription. Nature. 1981 Mar 26;290(5804):310-5.

Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E. TRANSFAC: transcriptional regulation, from patterns to profiles. Nucleic Acids Res. 2003 Jan 1;31(1):374-8.

Maynard J, Krawczak M, Upadhyaya M. Characterization and significance of nine novel mutations in exon 16 of the neurofibromatosis type 1 (NF1) gene. Hum Genet. 1997 May;99(5):674-6.

Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I. VISTA : visualizing global DNA sequence alignments of arbitrary length. Bioinformatics. 2000 Nov;16(11):1046-7.

McClelland M, Ivarie R. Asymmetrical distribution of CpG in an 'average' mammalian gene. Nucleic Acids Res. 1982 Dec 11;10(23):7865-77.

McLaughlin ME, Jacks T. Thinking beyond the tumor cell: Nf1 haploinsufficiency in the tumor environment. Cancer Cell. 2002 Jun;1(5):408-10. Review.

Messiaen LM, Callens T, Mortier G, Beysen D, Vandenbroucke I, Van Roy N, Speleman F, Paepe AD. Exhaustive mutation analysis of the NF1 gene allows identification of 95% of mutations and reveals a high frequency of unusual splicing defects. Hum Mutat. 2000;15(6):541-55.

Messiaen LM, Callens T, Roux KJ, Mortier GR, De Paepe A, Abramowicz M, Pericak-Vance MA, Vance JM, Wallace MR. Exon 10b of the NF1 gene represents a mutational hotspot and harbors a recurrent missense mutation Y489C associated with aberrant splicing. Genet Med. 1999 Sep-Oct;1(6):248-53.

Mills FC, Harindranath N, Mitchell M, Max EE. Enhancer complexes located downstream of both human immunoglobulin Calpha genes. J Exp Med. 1997 Sep 15;186(6):845-58.

Mitchell PJ, Timmons PM, Hebert JM, Rigby PW, Tjian R. Transcription factor AP-2 is expressed in neural crest cell lineages during mouse embryogenesis. Genes Dev. 1991 Jan;5(1):105-19.

Miyajima H, Miyaso H, Okumura M, Kurisu J, Imaizumi K. Identification of a *cis*-acting element for the regulation of SMN exon 7 splicing. J Biol Chem. 2002 Jun 28;277(26):23271-7.

National Institutes of Health Consensus Development Conference: Neurofibromatosis: conference statement. Arch. Neurol. 45: 575-578, 1988.

Ogbourne S, Antalis TM. Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes. Biochem J. 1998 Apr 1;331 (Pt 1):1-14.

Ohbayashi T, Shimada M, Nakadai T, Wada T, Handa H, Tamura T. Vertebrate TBP-like protein (TLP/TRF2/TLF) stimulates TATA-less terminal deoxynucleotidyl transferase promoters in a transient reporter assay, and TFIIA-binding capacity of TLP is required for this function. Nucleic Acids Res. 2003 Apr 15;31(8):2127-33.

Ohler U, Liao GC, Niemann H, Rubin GM. Computational analysis of core promoters in the Drosophila genome. Genome Biol. 2002;3(12):RESEARCH0087. Epub 2002 Dec 20.

Origone P, Defferrari R, Mazzocco K, Cunsolo CL, Bernardi BD, Tonini GP. Homozygous inactivation of NF1 gene in a patient with familial NF1 and disseminated neuroblastoma. Am J Med Genet. 2003 May 1;118A(4):309-13.

Otsuka, F.; Kawashima, T.; Imakado, S.; Usuki, Y.; Hon-mura, S.: Lisch nodules and skin manifestation in neurofibromatosis type 1. *Arch. Derm.* 137: 232-233, 2001.

Pickert L, Reuter I, Klawonn F, Wingender E. Transcription regulatory region analysis using signal detection and fuzzy clustering. Bioinformatics. 1998;14(3):244-51.

Praz V, Perier R, Bonnard C, Bucher P. The Eukaryotic Promoter Database, EPD: new entry types and links to gene expression data. Nucleic Acids Res. 2002 Jan 1;30(1):322-4.

Pugh BF. Control of gene expression through regulation of the TATA-binding protein. Gene. 2000 Sep 5;255(1):1-14.

Purandare S, Ota A, Neil S, Viskochil DH. Identification of *cis*-regulatory elements in the neurofibromatosis 1 gene. Am J Hum Genet. 1996 Jul;59(1):A157

Quandt, K., Frech, K., Karas, H., Wingender, E., Werner, T. MatInd and MatInspector - New fast and versatile tools for detection of consensus matches in nucleotide sequence data. Nucleic Acids Research 1995 23, 4878-4884.

Razin A, Riggs AD. DNA methylation and gene function. Science. 1980 Nov 7;210(4470):604-10.

Reuter, I. , (2000), Dissertation, http://www.biblio.tu-bs.de/ediss/data/20000317a/20000317a.html

Riccardi VM, Eichner JE. Neurofibromatosis: Phenotype, Natual History, and Pathogenesis, 2nd edition Balimore: Johns Hopkins University Press; 1986

Rodenhiser D, Zou MX, Groves TC, Butcher DT, Yee SP, Transcriptional regulation of the NF1 gene expression. Poster presented at the National Neurofibromatosis Foundation Consortium for the Molecular Biology of NF1 and NF2. Asoen, Columbia, June 2002.

Rodenhiser DI, Coulter-Mackie MB, Singh SM. Evidence of DNA methylation in the neurofibromatosis type 1 (NF1) gene region of 17q11.2. Hum Mol Genet. 1993 Apr;2(4):439-44.

Roulet E, Fisch I, Junier T, Bucher P, Mermod N. Evaluation of computer tools for the prediction of transcription factor binding sites on genomic DNA. In Silico Biol. 1998;1(1):21-8.

Santoro A, Tursz T, Mouridsen H, Verweij J, Steward W, Somers R, Buesa J, Casali P, Spooner D, Rankin E, et al. Doxorubicin versus CYVADIC versus doxorubicin plus ifosfamide in first-line treatment of advanced soft tissue sarcomas: a randomized study of the European Organization for Research and Treatment of Cancer Soft Tissue and Bone Sarcoma Group. J Clin Oncol. 1995 Jul;13(7):1537-45.

Sawada S, Florell S, Purandare SM, Ota M, Stephens K, Viskochil D. Identification of NF1 mutations in both alleles of a dermal neurofibroma. Nat Genet. 1996 Sep;14(1):110-2.

Scheffzek K, Ahmadian MR, Wiesmuller L, Kabsch W, Stege P, Schmitz F, Wittinghofer A. Structural analysis of the GAP-related domain from neurofibromin and its implications. EMBO J. 1998 Aug 3;17(15):4313-27.

Scherf M, Klingenhoff A, Werner T. Highly specific localization of promoter regions in large genomic sequences by PromoterInspector: a novel context analysis approach. J Mol Biol. 2000 Mar 31;297(3):599-606.

Schmidt MA, Michels VV, Dewald GW. Cases of neurofibromatosis with rearrangements of chromosome 17 involving band 17q11.2. Am. J. Med. Genet. 28:771-777.

Shen W, Liu DP, Liang CC. The regulatory network controlling beta-globin gene switching. Mol Biol Rep. 2001;28(3):175-83.

Sherman LS, Atit R, Rosenbaum T, Cox AD, Ratner N. Single cell Ras-GTP analysis reveals altered Ras activity in a subpopulation of neurofibroma Schwann cells but not fibroblasts. J Biol Chem. 2000 Sep 29;275(39):30740-5.

Side L, Taylor B, Cayouette M, Conner E, Thompson P, Luce M, Shannon K. Homozygous inactivation of the NF1 gene in bone marrow cells from children with neurofibromatosis type 1 and malignant myeloid disorders. N Engl J Med. 1997 Jun 12;336(24):1713-20.

Silverman ES, Le L, Baron RM, Hallock A, Hjoberg J, Shikanai T, Storm van's Gravesande K, Auron PE, Lu W. Cloning and functional analysis of the mouse 5-lipoxygenase promoter.Am J Respir Cell Mol Biol. 2002 Apr;26(4):475-83.

Skuse GR, Cappione AJ, French BL. A potential role for NF1 mRNA editing in the pathogenesis of NF1 tumors. Am J Hum Genet. 1997 Feb;60(2):305-12.

Smale ST. Transcription initiation from TATA-less promoters within eukaryotic protein-coding genes. Biochim Biophys Acta. 1997 Mar 20;1351(1-2):73-88.

Smit, AFA & Green, P RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html, 1996

Stark AM, Buhl R, Hugo HH, Mehdorn HM. Malignant peripheral nerve sheath tumours--report of 8 cases and review of the literature. Acta Neurochir (Wien). 2001;143(4):357-63; discussion 363-4.

Stocker KM, Baizer L, Coston T, Sherman L, Ciment G. Regulated expression of neurofibromin in migrating neural crest cells of avian embryos. J Neurobiol. 1995 Aug;27(4):535-52.

Stormo GD, Schneider TD, Gold L, Ehrenfeucht A. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in E. coli. Nucleic Acids Res. 1982 May 11;10(9):2997-3011.

Strachan T, Read A. Human Molecular Genetics 2. In: DNA structure and gene expression. Oxford, BIOS Scientific Publishers, 1999a, 1-26

Strachan T, Read A. Human Molecular Genetics 2. In: Human gene expression. Oxford, BIOS Scientific Publishers, 1999b, 169-208

Suzuki Y, Tsunoda T, Sese J, Taira H, Mizushima-Sugano J, Hata H, Ota T, Isogai T, Tanaka T, Nakamura Y, Suyama A, Sakaki Y, Morishita S, Okubo K, Sugano S. Identification and characterization of the potential promoter regions of 1031 kinds of human genes. Genome Res. 2001 May;11(5):677-84.

Szudek J, Joe H, Friedman JM. Analysis of intrafamilial phenotypic variation in neurofibromatosis 1 (NF1). Genet Epidemiol. 2002 Aug;23(2):150-64.

Tanaka K, Nakafuku M, Satoh T, Marshall MS, Gibbs JB, Matsumoto K, Kaziro Y, Toh-e A. S. cerevisiae genes IRA1 and IRA2 encode proteins that may be functionally equivalent to mammalian ras GTPase activating protein. Cell. 1990 Mar 9;60(5):803-7.

Tatusova TA, Madden TL. BLAST 2 Sequences, a new tool for comparing protein and nucleotide sequences. FEMS Microbiol Lett. 1999 May 15;174(2):247-50.

Thomas JE, Piepgras DG, Scheithauer B, Onofrio BM, Shives TC. Neurogenic tumors of the sciatic nerve. A clinicopathologic study of 35 cases. Mayo Clin Proc. 1983 Oct;58(10):640-7.

Thomas JW, Touchman JW. Vertebrate genome sequencing: building a backbone for comparative genomics. Trends Genet. 2002 Feb;18(2):104-8.

Thomson SA, Fishbein L, Wallace MR. NF1 mutations and molecular testing. J Child Neurol. 2002 Aug;17(8):555-61; discussion 571-2, 646-51.

Tinschert S, Naumann I, Stegmann E, Buske A, Kaufmann D, Thiel G, Jenne DE. Segmental neurofibromatosis is caused by somatic mutation of the neurofibromatosis type 1 (NF1) gene. Eur J Hum Genet. 2000 Jun;8(6):455-9.

Trahey M, McCormick F. A cytoplasmic protein stimulates normal N-ras p21 GTPase, but does not affect oncogenic mutants. Science. 1987 Oct 23;238(4826):542-5.

Upadhyaya M, Cooper DN: The mutational spectrum in neurofibromatosis 1 and its underlying mechisms. In: Upadhyaya M, Cooper DN editiors. Neurofibromatosis Type 1: From Genotype to Phenotype. Oxford, BIOS Scientific Publishers, 1998, 65-88.

Upadhyaya M, Ruggieri M, Maynard J, Osborn M, Hartog C, Mudd S, Penttinen M, Cordeiro I, Ponder M, Ponder BA, Krawczak M, Cooper DN. Gross deletions of the neurofibromatosis type 1 (NF1) gene are predominantly of maternal origin and commonly associated with a learning disability, dysmorphic features and developmental delay. Hum Genet. 1998 May;102(5):591-7.

Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. Nat Rev Genet. 2003 Apr;4(4):251-62.

van Tuinen P, Rich DC, Summers KM, Ledbetter DH. Regional mapping panel for human chromosome 17: application to neurofibromatosis type 1. Genomics. 1987 Dec;1(4):374-81.

Venkatesh B, Tay BH, Elgar G, Brenner S. Isolation, characterization and evolution of nine pufferfish (Fugu rubripes) actin genes. J Mol Biol. 1996 Jun 21;259(4):655-65.
Vijh S, Dayhoff DE, Wang CE, Imam Z, Ehrenberg PK, Michael NL. Transcription regulation of human chemokine receptor CCR3: evidence for a rare TATA-less promoter structure conserved between drosophila and humans. Genomics. 2002 Jul;80(1):86-95
Viskochil D, Buchberg AM, Xu G, Cawthon RM, Stevens J, Wolff RK, Culver M, Carey JC, Copeland NG, Jenkins NA, et al. Deletions and a translocation interrupt a cloned gene at the neurofibromatosis type 1 locus. Cell. 1990 Jul 13;62(1):187-92.
Viskochil D. Genetics of neurofibromatosis 1 and the NF1 gene. J Child Neurol. 2002 Aug;17(8):562-70; discussion 571-2, 646-51.
Viskochil DH. Gene Structure and Expression. In: Upadhyaya M, Cooper DN editiors. Neurofibromatosis Type 1: From Genotype to Phenotype. Oxford, BIOS Scientific Publishers, 1998, 65-88.
Viskochil DH. The Structure and Function of the *NF1* Gene: Molecular Pathophysiology. In: Friedman JM, Gutmann DH, MacCollin M, Riccardi VM editors. Neurofibromatosis: Phenotype, Natural History, and Pathogenesis. Baltimore and London: The Johns Hopkins University Press. 1999. pp. 119-141
Vogels A, Fryns JP. The Prader-Willi syndrome and the Angelman syndrome. Genet Couns. 2002;13(4):385-96.
Von Recklinghausen FD. Uever die multiplen fibrome der Hautund inhre beziehung zu den multiplen neuromen, Berlin: Hirschwald. 1982.
Wainer S. A child with axillary freckling and cafe au lait spots. CMAJ. 2002 Aug 6;167(3):282-3.
Wallace MR, Marchuk DA, Andersen LB, Letcher R, Odeh HM, Saulino AM, Fountain JW, Brereton A, Nicholson J, Mitchell AL, et al. Type 1 neurofibromatosis gene: identification of a large transcript disrupted in three NF1 patients. Science. 1990 Jul 13;249(4965):181-6. Erratum in: Science 1990 Dec 21;250(4988):1749.
Waterston RH, Lindblad-Toh K, Birney E, Rogers J, Abril JF, Agarwal P, Agarwala R, Ainscough R, Alexandersson M, An P, et al. Initial sequencing and comparative analysis of the mouse genome. Nature. 2002 Dec 5;420(6915):520-62.
White R, Nakamura Y, O'Connell P, Leppert M, Lalouel JM, Barker D, Goldgar D, Skolnick M, Carey J, Wallis CE, et al. Tightly linked markers for the neurofibromatosis type 1 gene. Genomics. 1987 Dec;1(4):364-7.
Whitmarsh AJ, Davis RJ. Transcription factor AP-1 regulation by mitogen-activated protein kinase signal transduction pathways. J Mol Med. 1996 Oct;74(10):589-607.
Wieczorek E, Brand M, Jacq X, Tora L. Function of TAF(II)-containing complex without TBP in transcription by RNA polymerase II. Nature. 1998 May 14;393(6681):187-91.
Wiestler OD, Radner H. Pathology of neurofibromatosis 1 and 2. In: Hudson SM, Hughes RAC, editors. The neurofibromatoses: a pathogenetic and clinical overview. Cambridge: Chapman and Hall; 1994. pp. 135-160
Wingender E, Chen X, Hehl R, Karas H, Liebich I, Matys V, Meinhardt T, Pruss M, Reuter I, Schacherer F. TRANSFAC: an integrated system for gene expression regulation. Nucleic Acids Res. 2000 Jan 1;28(1):316-9.
Wisdom R. AP-1: one switch for many signals. Exp Cell Res. 1999 Nov 25;253(1):180-5.
Wittinghofer A. Signal transduction via Ras. Biol Chem. 1998 Aug-Sep;379(8-9):933-7.
Wolffe AP, Jones PL, Wade PA. DNA demethylation. Proc Natl Acad Sci U S A. 1999 May 25;96(11):5894-6.

Xu GF, Lin B, Tanaka K, Dunn D, Wood D, Gesteland R, White R, Weiss R, Tamanoi F. The catalytic domain of the neurofibromatosis type 1 gene product stimulates ras GTPase and complements ira mutants of S. cerevisiae. Cell. 1990 Nov 16;63(4):835-41.

Xu GF, O'Connell P, Viskochil D, Cawthon R, Robertson M, Culver M, Dunn D, Stevens J, Gesteland R, White R, et al. The neurofibromatosis type 1 gene encodes a protein related to GAP. Cell. 1990 Aug 10;62(3):599-608.

Zanca A, Zanca A. Antique illustrations of neurofibromatosis. Int J Dermatol. 1980 Jan-Feb;19(1):55-8.

Zhang YY, Vik TA, Ryder JW, Srour EF, Jacks T, Shannon K, Clapp DW. Nf1 regulates hematopoietic progenitor cell growth and ras signaling in response to multiple cytokines. J Exp Med. 1998 Jun 1;187(11):1893-902.

Zhou T, Chiang CM. The intronless and TATA-less human TAF(II)55 gene contains a functional initiator and a downstream promoter element. J Biol Chem. 2001 Jul 6;276(27):25503-11. Epub 2001 May 04.

Appendix I Code for Frameslider

```
#!/usr/bin/perl

print "What size of the frame you will like to use, my majestry? ";  # Asking for the frame size to
look in the sequence
$size = <STDIN>;
chomp ($size);
print "What a good choice to use ", $size, " as the size!\n";

open_asequence ();
open_bsequence ();

$start = 1;                                                    # Set the value for human
starter
$position = 1;

while ($position <= 99999-$size) {
        $scounter = 0;
        while ($scounter < $size )      {
                if ($anucleo{$position} ne "-" ) {                          # check the counter to
be valid
                        $scounter = $scounter + 1;
                        if ($anucleo{$position} eq $bnucleo{$position}){
                                $a = $a + 1;
                        } else {
                                $a = $a;}
                } else {
                $scounter = $scounter;
                                                }
        $position = $position + 1;
        }

        open (STDOUT, ">>outfile.txt");
        write;                                                 #invoke format STDOUT to
STDOUT
        close (STDOUT) || die "can't close outfile: $!";


        $start = $start + 1;
        $position = $start;
        $a = 0;

}

sub open_asequence {
open (ASEQUENCE, "asequence.txt");
while ($anumber = <ASEQUENCE>) {
```

```
chomp ($anumber);
$anucleotide = <ASEQUENCE>;
chomp ($anucleotide);
$anucleo{$anumber} = $anucleotide;
}
close (ASEQUENCE);
}

sub open_bsequence {
open (BSEQUENCE, "bsequence.txt");
while ($bnumber = <BSEQUENCE>) {
chomp ($bnumber);
$bnucleotide = <BSEQUENCE>;
chomp ($bnucleotide);
$bnucleo{$bnumber} = $bnucleotide;
}
close (BSEQUENCE);
}


format STDOUT =
@<<<<<<<<< @<<<<<<<<<<<< @<<<<<<<<<< @<<<<<<<<<< @<<<<<<<
$start, $position - 1, $anucleo{$start}, $bnucleo{$start}, $a/$size;
.
format STDOUT_TOP =
Page @<<
$%

Start     End         ANucleotide BNucleotide Homology
========== ============ =========== =========== ========
.
```