Polymorphisms in the Flt1 gene and their relation to expression of the secreted Flt1 variant

Burouj Kayed Ajlouni

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University to fulfill the requirements for the degree of

Doctor of Philosophy in
Biomedical and Veterinary Sciences

William R. Huckle, Chairman
Klaus D. Elgert
Willard H. Eyestone
John L. Robertson
Jill C. Sible

September 7th, 2009
Blacksburg, Virginia

Keywords: Angiogenesis, Vascular Endothelial Growth Factor, Flt-1, sFlt-1, RNA Processing, SNPs, ESEs

Polymorphisms in the Flt1 gene and their relation to expression of the secreted Flt1 variant

Burouj Kayed Ajlouni

ABSTRACT

Vascular endothelial growth factor (VEGF) is a potent angiogenic agent. VEGF activates its biologic responses through two cell-surface receptors, Flt1 and Flk1. In addition to the transmembrane form of Flt1, the Flt1 gene also encodes a secreted, truncated form of the receptor (sFlt1) translated from an mRNA in which a portion of intron 13 is preserved. sFlt1 retains high affinity for VEGF and thereby inhibits its angiogenic activity. Intron 13 contains important *cis* elements involved in sFlt1 mRNA formation. Here, we test the hypothesis that polymorphisms in the human Flt1 gene, particularly SNPs at sites suspected to contain splicing or cleavage-polyadenylation signals, influence Flt1 pre-mRNA processing and rates of Flt1 and sFlt1 expression. The NCBI SNP database contained 23 SNPs in the region of interest, one each in exons 13 and 14. An independent human SNP screen confirmed several of the reported SNPs. The web-based ESEfinder software predicted that the exon 13/14 SNPs had reduced potential for recruitment of splicing components. To test effects of exonic SNPs on Flt1 pre-mRNA processing, wild type and mutant Flt1 minigene plasmids were transfected into NIH/3T3 cells. Both exonic SNPs were associated with ~40% decreases in Flt1:sFlt1 mRNA ratios determined by real-time PCR. To facilitate exploration of ESEs in regulated RNA splicing, a PERL computer program, “EXONerator,” was written to silence predicted ESEs without altering polypeptide sequence. These results support the notion that small changes in exon composition can have significant effects on splicing activity and underscore the utility of software tools for hypothesis generation.

Dedication

This manuscript is dedicated to my parents: my father, Kayed Ajlouni and mother, Muyasar Jabr. Without their unconditional love, support and encouragement, I would not be here today.

This manuscript is also dedicated to my five siblings: Kunouz, Mustafa and Mujahed, and to the memory of those who now dwell among the angels, Mukhtar and Ahmad.

Last but not least, this manuscript is dedicated to my husband, Mohannad Al-Saghir, who was there at every step with me on this journey with love and unwavering support. Thank you.

Acknowledgements

Funding support was provided by an assistantship provided by the Virginia Maryland Regional College of Veterinary Medicine. Additional grant support was provided by NIH grant R21 DK063098.

I would like to thank Dr. Carla Finkielstein and Dr. Linda Dahlgren for use of their lab space for some experimental steps. I would also like to thank the Virginia Tech Schiffert Health Center staff and the phlebotomists at Ohio University-Zanesville for their assistance in collecting blood samples. My sincere thanks also goes to all the volunteers who generously donated blood samples for human genomic DNA analysis.

I would also like to thank Jihane Najdi, Lee Falin and Dr. Brett Tyler for their input to The EXONerator. I would especially like to express my profound gratitude to Jihane Najdi, for lending her vast programming knowledge and expertise in PERL coding that eventually led to building and testing The EXONerator.

Additionally, I would like to thank Dr. Stephen Werre for his help with statistical analysis.

I would also like to thank each of my committee members, for all of their help and support during my graduate program.

Finally, I want to thank Dr. Huckle for his assistance and mentoring. I am sincerely blessed to have had the opportunity to work under his supervision. Without his extensive support and kind encouragement, none of this would have been possible.

Table of Contents

List of Tables and Figures

Chapter 1: Introduction and Literature Review

Physiological Angiogenesis

Angiogenesis and VEGF

In contrast to vaculogenesis (formation of new blood vessels from angioblastic precursor cells), angiogenesis is the formation of new blood vessels from pre-existing vasculature [5, 6]. Angiogenesis is an essential process, both in the embryo (for organ development and differentiation) and in the adult (for maintenance of vasculature, wound healing and reproductive cycle in females) [7, 8]. A healthy body controls this process by maintaining a balance of positive and negative regulatory factors, which are termed angiogenesis stimulating growth factors and angiogenesis inhibitors, respectively [3]. Angiogenesis is a multi-step process, whereby endothelial cells are activated and the basement membrane is degraded [8]. Cell migration then commences, as endothelial cells invade the extracellular matrix (ECM) and proliferate to form capillary lumen [8]. A reversal of the initiating processes (basement membrane reestablishes, cell proliferation ceases) allows for the stabilization of the new vascular network [8].

One of the most potent angiogenic stimulating factors is Vascular Endothelial Growth Factor (VEGF), an endothelial cell-specific mitogen that also mediates vascular permeability [9, 10]. There are multiple isoforms of VEGF (major species are 121, 145, 165, 185, 189 or 209 amino acids long), all encoded by a single gene with variants produced by alternative splicing [10-12]. $VEGF_{165}$, the most abundant form, (which will be referred to hereafter as VEGF for the purpose of this document) exists either in solution or bound to the cell surface and extracellular matrix by virtue of its affinity for heparin-like proteoglycans [11-13]. The essential role of VEGF in vascular development and homeostasis was demonstrated by several studies, whereby inactivation of a single VEGF allele in mice resulted in several developmental abnormalities, defective organ vascularization and ultimately, embryonic lethality between day 11 and day 12 [14, 15].

VEGF is involved in several stages of angiogenesis. It promotes the degradation of the ECM by mediating the secretion and activation of matrix metalloproteinase enzymes [8]. VEGF also induces normally quiescent endothelial cells to divide, migrate

to sites of neovascularization and evade cell death via expression of antiapoptotic proteins [8, 16]. VEGF may act in increasing vascular permeability (which is why it is also known by the name of Vascular Permeability Factor or VPF) and inducing monocyte migration [7, 8].

VEGF gene expression is upregulated by several different factors, such as hypoxia [17], hormones including estrogen [18], nitric oxide [19] and several growth factors including transforming growth factor beta (TFG- ß), tumor necrosis factor alpha, epidermal growth factor, keratinocyte growth factor, insulin-like growth factor 1, fibroblast growth factor and platelet-derived growth factor [20-22].

VEGF Receptors

VEGF mediates its actions via its receptor molecules, two cell-surface, membrane-spanning tyrosine kinases, the Flt1 receptor (or VEGFR-1) and KDR (or VEGFR-2/murine Flk1). Flt1 and KDR are expressed principally in endothelial cells, both quiescent and proliferating [23-26]. The receptors have similar protein structures and share a 43.2% identity in their amino acid sequences [27]. The structure of both receptors is made up of seven immunoglobin-like (Ig-like) loops (the extracellular ligand-binding region), a single transmembrane domain, a juxtamembrane domain of 60-70 amino acid residues and a tyrosine kinase domain which is interrupted by a kinase insert domain (Figure 1.1, top) [25, 27, 28].

While Flt1 has a much higher (at least 10-fold) affinity for VEGF than does KDR, it has a 10-fold weaker tyrosine kinase activity, which points to differing roles of the receptors [23, 29]. Flt1 is essential for endothelial cell morphogenesis, while KDR mediates mitogenesis [30]. Hypoxia upregulates expression of Flt1 and, in some cases, KDR (as it does with VEGF) [8, 31, 32]. The VEGF receptor levels are also upregulated secondary to VEGF stimulation, which ultimately increases the signal strength of VEGF [8]. However, other factors that stimulate overexpression of VEGF have been shown to have variable effects on VEGF receptor expression. For example, TGF-ß, an upregulator of VEGF, has been shown to downregulate the expression of VEGF receptors [8].

Binding of VEGF to its receptors leads to their dimerization, autophosphorylation, and activation of signal transduction pathways, which results in the cellular responses noted above [33]. Flt1 is essential for the normal formation and organization of blood vessels, particularly in embryonic development, but its exact function in adult vasculature is not well understood [23, 34, 35]. The first three Ig-like domains of Flt1 are essential for VEGF binding, upon which the Flt1 receptor becomes phosporylated [36, 37]. The fourth Ig-like loop is essential in the subsequent dimerization of Flt1, whereupon the Flt1 receptor is activated and able to carry out VEGF-mediated functions [36, 38].

sFlt1

The Flt1 gene produces two forms of the Flt1 protein via alternative processing of the pre-mRNA transcript: a full length membrane bound Flt1 receptor and a soluble truncated Flt1 receptor, known as sFlt1 (Figure 1.1) [39]. An interesting observation was the high amino acid conservation between the C-terminal regions of mouse sFlt1 and human sFlt1 (87%), as opposed to the lower conservation of the N-terminal regions of Flt1 in human and mouse (78%) [39]. This strong evolutionary conservation points to the biological significance of sFlt1 in mammals [39].

Full length Flt1 is formed via splicing of all 30 exons [26]. On the other hand, sFlt1 is produced as an alternate transcript, whereby a portion of intron 13 is retained, and exons 14-30 are deleted [2, 26]. The resulting sFlt1 protein is made up of the first 6 Ig-like domains (where VEGF binding occurs) but lacks the 7th Ig-like domain, the transmembrane domain and the intracellular tyrosine kinase domains of Flt1 [26]. Instead, sFlt1 exhibits a unique C-terminus made up of 31 amino acids (corresponding to the retained intron 13 portion) that is not found in Flt1 [39]. sFlt1 retains the high VEGF binding affinity of Flt1, but unlike the full length form, sFlt1 serves as an inhibitor for VEGF [2, 26]. sFlt1 is thought to act in the manner of a “decoy” receptor, as it binds VEGF with high affinity, thus decreasing the availability of VEGF to the functional receptors, and preventing mitogenic activity of VEGF to proceed [8]. It may also act in a dominant negative fashion, by forming nonfunctional heterodimers with membrane

bound VEGF receptors [32]. This “decoy” function of sFlt1 appears to allow for regulating the rate of angiogenesis in the healthy body.

In addition to regulation of VEGF-mediated angiogenesis, sFlt1 also appears to prevent blood vessel growth from occurring as it is naturally expressed at high levels in several tissues where a lack of vascularization is essential. This includes the cornea, the enthothelium and in placental trophoblasts (where it is essential to avoid maternal and fetal blood mixing) [40]. This suggests that sFlt1 is effective at naturally inhibiting angiogenesis in vivo, and can be manipulated to inhibit angiogenesis at other sites in the body.

When sFlt1 is overexpressed, it tends to be associated with disease. In pregnant women, overexpression of sFlt1 is associated with preeclampsia which – by blocking VEGF function –appears to directly cause hypertension and renal dysfunction [41]. In fact, it has recently been suggested that sFlt1 levels can be an important prognostic factor, as pregnant women with high sFlt1 levels in the first trimester tend to develop late-onset preeclampsia [42]. Lower expression of sFlt1 can also be a problem; it has recently been suggested that there may also be a potential link between lower expression sFlt1 levels and ovarian hyperstimulation syndrome [43], and low sFlt1 levels and acute mountain sickness [44]

The relative levels of Flt1 and sFlt1 mRNAs vary according to physiological conditions, which implies a regulatory mechanism whereby the level of angiogenesis is regulated by the ratio of the sFlt1 to Flt1 [45-47]. Previous work in our laboratory suggests that levels of sFlt1 and Flt1 are normally in a regulated ratio, whereby an increase of one type of pre-mRNA processing would result in a reciprocal decrease in the other [2]. It has been suggested that there are (as yet not understood) regulatory mechanisms in place that control the processes of forming Flt1 (via intron splicing) or sFlt1 (via cleavage-polyadenylation within intron 13), forming a carefully controlled ratio between the two products and thus regulating the bioavailability of VEGF as appropriate [2].

Pathological Angiogenesis

Aberrant angiogenesis plays a key role in the development and progression of many pathological conditions. Diseased conditions can be linked to deficient angiogenesis (such as impaired wound healing and impaired collateral vessel formation in diabetes[48]) or excessive angiogenesis (such as arthritis[49] and cancer [50]).

Over the past decade, much attention has been directed at research on tumor angiogenesis [8]. Despite more than 50 years of research and more than $2.5 billion distributed to researchers (1946-2003) by the American Cancer Society alone, there is still no cure for cancer. According to the American Cancer Society (www.cancer.org), cancer is the second-leading cause of death in the US (accounting for almost one-quarter of deaths). It is estimated that in 2007, 1.4 million people will be diagnosed with cancer. One in three American women and one in two American men will develop cancer over their lifetime. A better understanding the role of angiogenesis in cancer will give researchers a valuable opportunity to manipulate this process to produce more effective therapies for cancer.

The progression of a solid tumor must be accompanied by development of a vascular supply to support the continued nutrient demands of the growing mass [50]. Tumor growth is accompanied by an expansion of its vascular network, which also helps in facilitating tumor metastasis [8]. A tumor can survive without a vascular supply (by absorbing nutrients and oxygen via simple diffusion) only up to a size of 1-2 mm^3 [8]. To grow any further, the tumor must establish a vascular supply that infiltrates the tumor mass [51]. However, while the tumor can efficiently produce a vascular supply, the blood vessels it produces are different from normal blood vessels, as they are often unorganized, irregular, and highly permeable [52]. This not only has growth implications (uneven delivery of nutrients to the tumor, so growth can be uneven) but also therapeutic implications as therapies delivered via the bloodstream (such as chemotherapy) may not be delivered efficiently to all parts of the tumor [52]. As a result, targeting the vasculature of the tumor, rather than the tumor cells, is an increasingly

attractive target for therapy [52]. This approach, termed antiangiogenic therapy, has several advantages, the most important of which being the vasculature instability (especially when compared to the high resilience of neoplastic cells) [52]. Antiangiogenic therapy is also attractive because a pro-angiogenic tumor in cancer patients generally indicates a poor prognosis, partly owing to the strong metastatic potential from the dense vascular supply surrounding the tumor [53].

Antiangiogeneic therapy is still a relatively new field, but there have been many advances to date. Research focus has included natural inhibitors (such as interleukin 12 and interferon α) as well as decoys, tyrosine kinase inhibitors, soluble receptors and synthetic compounds targeting pro-angiogeneic factors such as VEGF [53]. Perhaps the best known antiangiogenic therapy is bevacizumab (marketed as “Avastin” by Genetech), which was approved by the FDA for treating metastatic colorectal cancer in February, 2004 [54]. Bevacizumab is a humanized murine anti-VEGF monoclonal antibody that inhibits the growth of human tumor cell lines in nude mice [54]. Avastin is highly successful when used in combination therapies, and five years later, is now being used for three additional cancers: Non-Squamous Non-Small Cell Lung Cancer, metastatic breast cancer and metastatic renal cell carcinoma [45].

Another significant antiangiogenic drug is PTK787 (marketed as “Vatalanib” by Novartis). This oral inhibitor of the tyrosine kinase activity of the VEGF receptors shows significant inhibition of tumor growth and inhibition of metastasis in preclinical studies on a variety of tumors [33]. PTK787 is currently being evaluated in seven Phase III clinical trials for several different types of cancers [55]. CEP-7055 (marketed by Cephalon) is another oral inhibitor of VEGF receptors, and shows potent antiangiogenic activity in tumors transplanted into mice [56]. CEP-7055 recently successfully completed Phase II clinical trials for treatment of glioma and colon carcinoma [33, 57].

Another approach for therapy is exemplified by AVE-500 (marketed as “the VEGF-Trap” or “Aflibercept” by Sanofi-Aventis/Regeneron), which acts as a soluble decoy receptor for VEGF in a manner similar to sFlt1 [33]. Treatment of animal tumors showed potent antitumor activities, even in advanced phases of tumor development [58]. AVE-500 is currently being studied in an international Phase II clinical trial on recurrent epithelial ovarian cancer [59].

Antiangiogenic therapies are effective because VEGF is expressed at high levels by many human tumors, including lung [60], breast [61], gastrointestinal tract [62], kidney [63], bladder [63], ovary [64] and glioblastoma [65]. Many tumor cell lines also secrete VEGF at high levels *in vitro* [66]. VEGF overexpression in tumors is further amplified by upregulation of VEGF due to hypoxia, a characteristic of most tumors, due to the poorly organized vascular structure [8]. In addition, since VEGF binding to receptors can result in upregulated expression of VEGF receptors, the VEGF-mediated response is further amplified in these tumors [8]. VEGF receptor expression occurs in several types of cancer cells, including leukemic cells, Kaposi's sarcoma, breast cancer cells and colorectal carcinoma cells [67, 68].

Blocking VEGF activity through the use of neutralizing antibodies for VEGF (thereby blocking receptor activity) can inhibit tumor growth *in vivo* [51]. Because of the important role of VEGF (and the VEGF receptors through which it mediates its biological response) in the growth of tumors, blocking its effect has become a main focus of antiangiogenic therapy [52]. Interestingly, overexpression of sFlt1 inhibits the growth of tumors [32]. In addition, transient transfection of tumor cells lines with an sFlt1 gene resulted in a significant inhibition of implantation and subsequent growth of tumors when implanted into nude mice [32, 69]. These findings suggest that therapies that manipulate native sFlt1 expression may be an important therapy approach. Before researchers are able to produce such therapies, the mechanisms controlling the regulation of Flt1 vs. sFlt1 expression must first be understood so as to be able to manipulate them rationally.

Mechanisms of pre-mRNA processing

Although only a small portion of the human genome appears to code for proteins, the phenomenon of pre-mRNA processing can serve to generate multiple different protein variants from the same gene, thus increasing the variety of gene products derived from a set number of genes [70]. The gene product that is produced can depends partly on the sequence of the gene, as well as other factors such as the cell

type/location/fate and signals from surrounding cellular environment, all of which help determine the protein variant to be produced [70]. In addition, cellular events can determine stability and half-life of the mRNA transcripts, allowing for a cell-specific regulation of protein products [70].

The term RNA processing is used to describe a series of biochemical events that result in the maturation of the RNA molecules [71]. A protein-encoding RNA transcript goes through three main stages: initially during transcription, it is a “nascent transcript.” Upon completion of transcription, it is a “primary transcript,” and after post-transcriptional modifications are complete, it becomes a “mature transcript” that is now ready to be exported from the nucleus and translated [71]. Pre-mRNA undergoes extensive processing to produce the mature transcript, both at the ends (5’ capping and 3’cleavage/polyadenylation) and internally in the transcript itself (splicing) [71]

In the nucleus, the pre-mRNA is associated with several proteins, which make up the heterogenous ribonucleoproteins (hnRNPs). These proteins are thought to act in a similar manner as the single stranded DNA-binding proteins in transcription; that is, they ensure that the RNA is single stranded and free of any secondary structure, and they facilitate the interactions of RNA with other components of the splicing machinery (and in some cases, as in splicing proteins, acting as specific binding sites for RNA processing factors) [71].

Although the mechanisms of mRNA processing are usually studied as if they occur separate from transcription, there is irrefutable evidence that the two process are intricately linked [71-74]. During the initiation of transcription, a collection of proteins called transcription factors mediate the binding of RNA polymerase [74]. Cleavage polyadenylation stimulatory factor (CPSF) associates with general transcription factors, which are recruited alongside RNA polymerase II (Pol II) [74]. The Pol II C terminal domain (CTD) is highly phosporylated during the process of elongation, and elongation factors replace the general transcription factors [74]. These elongation factors move alongside the Pol II elongation complex, and are thus able to process the transcript as it is released from the Pol II complex [74].

The pre-mRNA transcript undergoes three main modifications in the cell nucleus: 5' capping, RNA splicing and 3' cleavage/polyadenylation. The first modification to

occur, 5’ capping (addition of 7-methylguanosine to 5' end of mRNA), involves a rapid reaction, catalyzed by three enzymes: mRNA guanyltransferase, a phosphatase and a methylase [71, 74]. The reaction involves the loss of the terminal phosphate groups, and the guanosine diphosphate residue is then methylated, resulting in the inverted guanosine triphosphate residue that ‘caps’ the 5’ end[71]. The addition of the mRNA cap occurs before the transcript exceeds ~30 nucleotides in length, and is thought to indicate the point at which transcription initiation ends, and elongation begins [71, 74]. This cap appears to have several functions: it is essential for export of the mRNA from the nucleus, aids in the binding of the mRNA to ribosomes (essential for translation) and helps protect the mRNA molecule from degradation in the nucleus [71, 75].

The second modification is RNA splicing, which involves the removal of introns from the pre-mRNA transcript [71]. In mammals, more than 90% of most genes are made up of introns, and splicing of these noncoding introns is essential for correct translation of the mRNA into a functional protein [71, 74].

The spliceosome (a complex of specialized RNA and protein subunits that facilitates the removal of introns) attaches to the mRNA transcript at sites termed “splice signals”, which include the 5’ splice site (typically a GU nucleotide sequence adjacent to the 3’ end of the preceding exon), the 3’ splice site (an AG sequence adjacent to the 5’ end of the following exon), the branch point (found within the intron, 20–40 nucleotides upstream of the 3’ splice site and contains the conserved A required for the first step of splicing) and the polypyrimidine tract (also within the intron) (Figure 1.2) [76]. In addition, exonic sequences such as Exonic Splice Enhancers (ESEs) can affect splicing site, as SR proteins recruited to ESE sites have been shown to interact with the CTD during RNA processing [74, 77]. While the details of cotranscriptional splicing remain to be elucidated, it is clear that CTD interaction (via a specific CTD-binding domain) with splicing factors is essential for RNA processing[74].

The last step in pre-mRNA processing is cleavage/polyadneylation, which leads to the mature mRNA transcript with the polyadenylate tail added to the cleaved 3' end of mRNA [71-73]. Cleavage of the 3’ end occurs at a sequence (usually ‘CA’) about 10-30 nucleotides downstream of the cleavage/polyadneylation signal (AAUAAA sequence) located near the 3’ end of the pre-mRNA [78]. Another signal (a GU-rich sequence) lies

about 30 nucleotides downstream of the site of cleavage [78]. The processes of cleavage and polyadenylation are linked, and it is not clear how the cleavage complex transitions to the polyadneylation complex [78]. Cleavage stimulating factor (CstF) binds to the CTD domain at the downstream signal for cleavage/polyadenylation and recruits other cleavage and polyadneylation factors [71, 78]. CstF is involved in the cleavage reaction, but not in polyadenylation. Cleavage polyadenylation specificity factor (or CPSF) is a multi-subunit complex that recognizes and binds directly to the cleavage/polyadenylation sequence, and is involved in both cleavage and polyadneylation reactions [71, 78]. After binding of CstF and CPSF, a protein complex forms, made up of cleavage factors I and II (both required for cleavage, but not for polyadneylation) and the polyadenylate polymerase (PAP) enzyme which facilitates the cleavage of the RNA [78]. PAP is also involved in polyadenylation, as it catalyzes the addition of the 200 A residues to the 3' end of the RNA molecule [78]. The polyadenylate binding protein (PABP) serves to increase productivity of PAP after a short oligo-tail has been generated, and in involved in extending the poly-A tail as well as controlling its length [71, 78]. Polyadenylation is important in stabilizing the mRNA transcript and protecting the transcript from decay [71, 75]. The poly-A tail is also important in the initiation of translation [78].

Alternative splicing, a common process in gene regulation, is a process whereby alterations in splice site choices determine the portions of coding sequence which will be retained in the final mRNA transcript [79, 80]. This process allows one gene to produce different forms of the same protein that differ in the final peptide sequence and thus in their biological function [79, 81]. Alternative splicing may retain specific exon and even intron sequences, by using alternate splicing sites [72, 80].

The Flt1 gene produces full length Flt1 and truncated sFlt1 via alternative processing [39]. The levels of Flt1 and sFlt1 mRNAs vary according to physiological conditions, but the process of regulation is still poorly understood [45-47]. Previous work in this laboratory has shown that there are several important regulatory regions in intron 13 of the mouse Flt1 gene that can determine whether pre-mRNA is processed to form Flt1 (via intron splicing) or sFlt1 (via cleavage-polyadenylation) (Figure 1.3).

A series of experiments was performed in which one or more of these cleavage-polyadenylation signals were deleted [2]. When all six signals were deleted, sFlt1 levels decreased to extremely low levels [2]. However, deletion of any combination of signals #1-#4 did not appear to have a significant effect of sFlt1 levels, indicating that signals #5 and #6 were capable of producing sFlt1 autonomously [2]. Further support of this idea came from experiments involving deletion of signals #5 and/or #6, which reduced sFlt1 protein expression by 77% [2]. When signals #1 and/or #2 were deleted along with #5 and/or #6 showed no further decrease in sFlt1 levels, but if signal #3 was the one deleted instead of signals #1/#2, it resulted in a further decrease to levels where sFlt1 was no longer detectable [2]. Thus, it appears that multiple signals have the ability to allow for sFlt1 expression, with no one single signal being singularly responsible for sFlt1 expression [2]. It is interesting to note that in these experiments, a decrease in sFlt1 mRNA levels was accompanied by a reciprocal increase in Flt1 mRNA levels, which suggests that lack of sFlt1-producing elements allow for Flt1 processing instead [2]. There may be regulatory mechanisms in place that control the ratio of sFlt1 and Flt1 [2].

Previous work in the lab has shown that single base pair changes at key regulatory sites, such as the 5' splice site, can dramatically alter the apparent activity of signals related to splicing of intron 13 [107]. In the current work, we are interested in finding any functional effects of naturally occurring single base pair changes in the form of single nucleotide polymorphisms (or SNPs) found in intron 13 as well as the surrounding exons (exons 13 and 14). In addition, the Flt1 gene is highly conserved across species, and we are interested in finding the areas that are conserved and correlating these conserved areas with a functional reason as related to production of Flt1 protein variants. A more thorough understanding of how Flt1 protein variants are produced is important because of the natural inhibitory effect of sFlt1 on angiogenesis and the potential for adaptation of this information for use in anti-angiogenic therapies in cancer.

Hypothesis

Hypothesis: Naturally-occurring polymorphisms in key regulatory regions of the Flt1 locus can alter the relative expression of FtI1 and sFlt1 mRNA and protein.

Specific Aims

1. To identify SNPs in the human Flt1 gene in regions likely to influence sFlt1 expression, by querying public SNP databases and screening additional human genomic DNAs.

2. To identify evolutionarily conserved regions in the Flt1 gene that may affect sFlt1 expression.

3. To test the effects of genetic polymorphisms identified in Aims 1 and 2 on Flt1 pre-mRNA processing, using a human Flt1 minigene expression system.

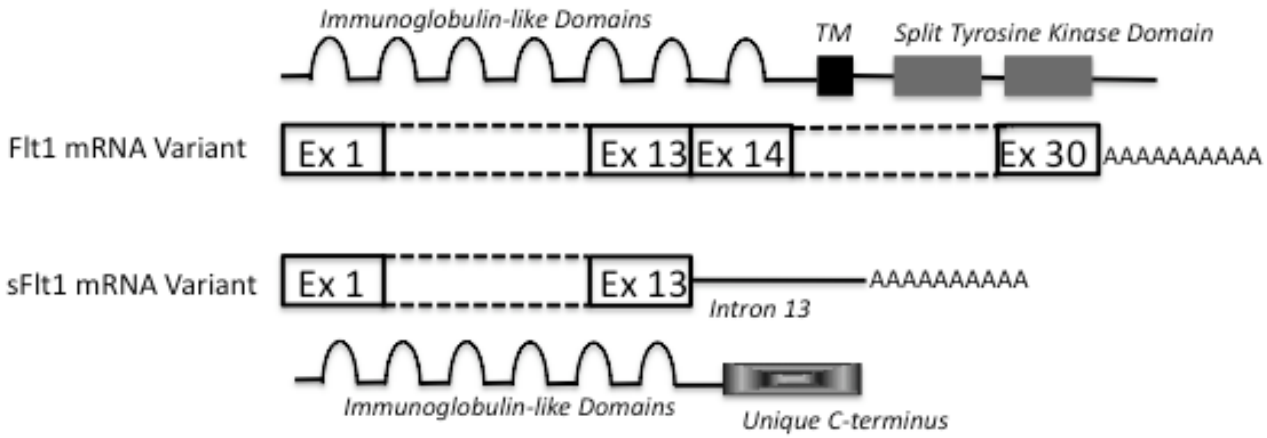


Figure 1.1: Flt1 and sFlt1 mRNA variants. The Flt1 gene produces two forms of the Flt1 mRNA via alternative processing of the pre-mRNA transcript: a transcript that codes for the full length membrane bound Flt1 receptor (top) and a transcript that codes for the soluble truncated Flt1 receptor, known as sFlt1 (bottom). Full length Flt1 is formed via splicing of all 30 exons, while sFlt1 is produced as an alternate transcript, whereby a portion of intron 13 is retained, and exons 14-30 are deleted. The resulting sFlt1 protein is made up of the first 6 Ig-like domains (where VEGF binding occurs, thus retaining full binding affinity for VEGF) but lacks the 7th Ig-like domain, the transmembrane domain and the intracellular tyrosine kinase domains of Flt1. Instead, sFlt1 exhibits a unique C-terminus made up of 31 amino acids (corresponding to the retained intron 13 portion) that is not found in Flt1. TM: transmembrane domain, Ex: exon. Figure modified from [2].

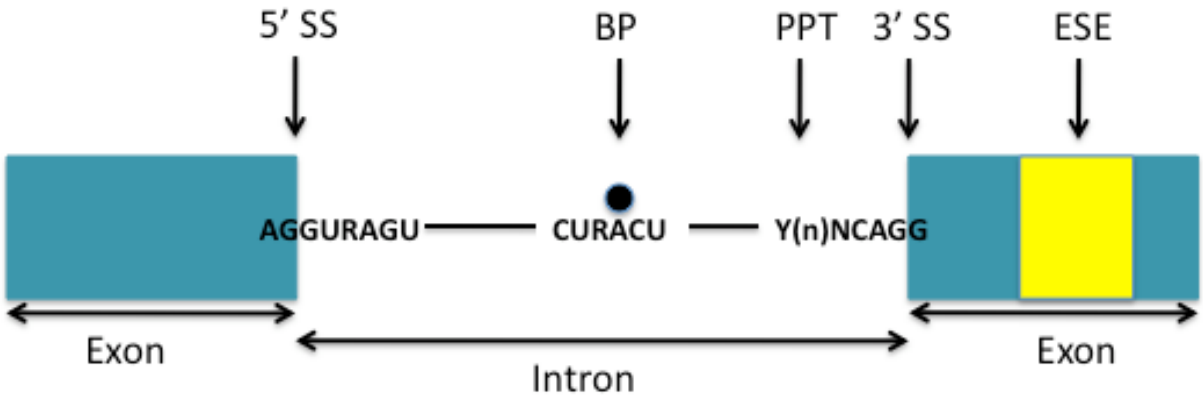


Figure 1.2: Pre-mRNA "Splice Signals". During RNA processing, the spliceosome attaches to the mRNA transcript at sites termed "splice signals", which include the 5' splice site (5'SS, typically a GU nucleotide sequence adjacent to the 3' end of the preceding exon), the 3' splice site (3'SS, an AG sequence adjacent to the 5' end of the following exon), the branch point (BP, found within the intron, 20–40 nucleotides upstream of the 3' splice site and contains the conserved A required for the first step of splicing) and the polypyrimidine tract (PPT, also within the intron). In addition, exonic sequences such as Exonic Splice Enhancers (ESEs) can affect splicing site. (R: purine, Y: pyrimidine). Figure modified from [4].

Figure 1.3: Mouse Flt1 Intron 13 Consensus Cleavage/Polyadneylation Signals. Previous work in this laboratory [2] has identified six cleavage/polyadenylation signals (annotated in figure) in intron 13 of the mouse Flt1 gene that can determine whether pre-mRNA is processed to form Flt1 (via intron splicing) or sFlt1 (via cleavage-polyadenylation). A series of experiments was performed in which one or more of these cleavage-polyadenylation signals were deleted. When all six signals were deleted, sFlt1 levels decreased to extremely low levels. However, deletion of any combination of signals #1-#4 did not appear to have a significant effect of sFlt1 levels, indicating that signals #5 and #6 were capable of producing sFlt1 autonomously. This was also indicated by an experiment where deletion of signals #5 and/or #6 reduced sFlt1 protein expression by 77%. When signals #1 and/or #2 were deleted along with #5 and/or #6 showed no further decrease in sFlt1 levels, but if signal #3 was the one deleted instead of signals #1/#2, it resulted in a further decrease to levels where sFlt1 was no longer detectable. Thus, it appears that multiple signals have the ability to allow for sFlt1 expression, with no one single signal being singularly responsible for sFlt1 expression. Figure modified from [2].

Chapter 2: Effects of Known Exonic Single Nucleotide Polymorphisms (SNPs) on Flt1 and sFlt1 expression

Abstract

Vascular endothelial growth factor (VEGF) is a potent angiogenic agent. VEGF activates its biologic responses through two cell-surface receptors, Flt1 and Flk1. In addition to the transmembrane form of Flt1, the Flt1 gene also encodes a secreted, truncated form of the receptor (sFlt1) translated from an mRNA in which a portion of intron 13 is preserved. sFlt1 retains high affinity for VEGF and thereby inhibits its angiogenic activity. Intron 13 contains important *cis* elements involved in sFlt1 mRNA formation. Here, we test the hypothesis that polymorphisms in the human Flt1 gene, particularly SNPs at sites suspected to contain splicing or cleavage-polyadenylation signals, influence Flt1 pre-mRNA processing and rates of Flt1 and sFlt1 expression. The NCBI SNP database contained 23 SNPs in the region of interest, one each in exons 13 and 14. To test effects of exonic SNPs on Flt1 pre-mRNA processing, wild type and mutant Flt1 minigene plasmids were transfected into NIH/3T3 cells. Both exonic SNPs were associated with ~40% decreases in Flt1:sFlt1 mRNA ratios determined by real-time PCR. Testing effects of silenced predicted ESEs in the areas of each SNP within the same model (mutant Flt1 minigene plasmids) revealed inconclusive results, possibly due in part to incomplete silencing of ESEs in region. These results support the notion that small changes in exon composition can have significant effects on splicing activity.

Introduction to SNPs

SNPs and Disease

When comparing two randomly chosen human genomes, there is only 0.1% variation in sequence, with the remaining 99.9% of the genome identical [82]. Variations in the genome affect individuals in several ways, including susceptibility to disease onset and progression and even responsiveness to therapy [59]. One example of such variation is the single nucleotide polymorphism (SNP): single base pair differences in the sequence of individual genomes occurring at a frequency of 1% or greater in at least one distinct population [83, 84]. In addition to being the simplest form of variation of DNA sequence, SNPs are also the most frequent and thus are crucial to probing and understanding the genetic basis of disease [85]. SNPs play a key role in several diseases [86, 87], including asthma [88], Crohn's disease [89], hypertension [86], Alzheimer's disease [90], Type I [91] and Type II [92] diabetes, arthritis (juvenile idiopathic arthritis [93], rheumatoid arthritis [94] and osteoarthritis [95]), obesity [96], bipolar disorder [97] and cancer (examples are metastatic breast cancer [98], colorectal cancer [99], familial prostate cancer [100] and lung cancer [101]). In addition, a rare functional SNP in the Flt1 promoter region observed in colon, breast and lung tumor cell lines (with an incidence of ~6% in Caucasian individuals), was observed to be specifically responsive to the p53 transcriptional regulator (a tumor suppressor gene) [102]. The study of SNPs has great potential for generating advances in therapeutics. For example, if one were to identify SNPs involved in a specific disease that result in a varied response to therapy, the study of such SNPs could lead to individually tailored and highly efficient therapies based on genome sequence [87].

Possible Effects of SNPs

There are multiple ways in which a SNP can affect the fate of an mRNA transcript or protein, depending on the type of SNP and its location in the genome. SNPs occur in the genome as transitions (interchanges of purines or of pyrimdines) or

transversions (interchanges between purine and pyrmidine bases), with an estimated frequency of 1 SNP per 1000 bp [87]. A screen of SNPs identified an approximately equal distribution of SNPs in the noncoding and coding regions of the genome; of the SNPs occurring in coding regions, 50% were synonymous SNPs (silent SNPs that did not result in an amino acid change) while the other 50% were nonsynonymous (SNPs resulting in a change in amino acid sequence) [86]. Both synonymous and nonsynonymous SNPs have the potential to influence mRNA processing [86]. Such changes, as well as changes to protein function (due to changes in structure), may result in variation between individuals in disease susceptibility, severity and progression, as well as response to therapies [87].

SNPs in the coding region of a gene can result in changes to the encoded product, such as single amino acid substitutions, truncated product (due to introduction of an early stop codon) or a mis-sense protein (due to an insertion or deletion SNP that results in a frameshift effect) [12]. Such SNPs are easier to detect and study, but they are only a subset of currently identified SNPs [12]. SNPs can also occur in regulatory regions, such as within introns, or promoters, and these SNPs could affect processes such as splicing efficiency and stability of the mRNA transcript. SNPs can affect splicing by their presence in various locations on the sequence, such as canonical splice site consensus sequences, exonic splice enhancers (ESEs) and intronic splice enhancers (ISE) [103]. SNPs that affect splicing are of particular interest in the present case, since splicing efficiency may affect the relative abundance levels of full length Flt1 (splicing of introns) or the truncated sFlt1 (cleavage and polyadenylation) [2].

SNPs and Flt1

The garnered interest in SNPs and their potential medicinal importance has led to large-scale SNP genotyping in recent years, which has resulted in the creation of several databases that document SNP locations in genes. The most comprehensive of these is the NCBI SNP database [3, 85]. Another database used in this study is that accessed via the SNPBrowser program (v 5.1), a collection provided by Applied Biosystems which contains over 160,000 SNPs as well as the ~3.2 million SNPs in the

NCBI database [104]. This vast amount of SNP information allows for studying the location and functional effects of SNPs in a way not before possible [85].

For the purpose of the present study, we chose to look at the Exon13- Intron 13-Exon-14 of the Flt1 gene, a sub-locus that our laboratory has shown to be important in determining the outcome of mRNA processing. In addition, we included consideration of short regions from the flanking introns 12 and 14, owing to their potential influence on the recognition of exons 13 and 14 as exons by the RNA processing machinery [2]. The total length of the Flt1 region screened was 9185 bp, an area that likely plays an important part in the processing of pre-mRNA transcripts to mRNAs encoding Flt1 and sFlt1. We have hypothesized that different variants of SNPs in these regions may lead to differential processing of pre-mRNA transcripts that contain them.

It is estimated that about 15% of all point mutations resulting in a genetic disease do so by affecting splicing, via direct inactivation of an essential splice site, introduction/activation of a cryptic splice site or by hindering the activity of splicing regulatory elements [105]. The effect of base pair changes in exons can be obvious when they affect the sequence of amino acid, but even if the changes result in no change in protein product, there can still be an effect on splicing [105]. One such change is that which affects the exonic splice enhancers (ESEs), short sequences found within an exon that serve as binding sites for a family of highly conserved splicing factors called the SR proteins [4, 77]. These SR proteins play several roles in the splicing pathway, promoting the definition of exon boundaries by directly recruiting splicing machinery and/or repressing the effects of surrounding silencer elements [77]. Disruption of the ESE sequence (such as by presence of a SNP within the ESE sequence) can affect the strength of binding of these proteins, which may lead to the exon being fully or partially “skipped” during the splicing process, thus modifying the final protein product [77][28]. This disruption of ESE sequences is of particular interest in the case of the Flt1 gene, as alternative splicing leads to two distinct protein products of the Flt1 receptor.

Materials and Methods

Screening Flt1 Region for SNPs

At the beginning of the study in 2006, a comprehensive survey of the NCBI database and SNP Browser for the 9185 bp Flt1 region of interest revealed thirty-six SNPs (Table 2.1). In each of Exons 13 and 14, only one SNP was found, both of which are silent with respect to predicted amino acids sequence. Both exonic SNPs were described as "double hits" (D), which is defined by NCBI as having been "validated by frequency or genotype data with minor alleles observed in at least two chromosomes." In the intronic regions, a total of 4 SNPs had been reported in the studied portion of Intron 12, 17 SNPs in Intron 13 and 13 SNPs in the Intron 14 fragment. As SNP discovery remains a work in progress, most of the SNPs did not have any population data included, and the relevance of these SNPs to Flt1 expression is difficult to interpret. This lack of population data highlights the fact that current SNP data is incomplete and thus the need for a more thorough SNP screen in the Flt1 region.

Exonic SNPs and ESEs

Due to the incomplete nature of the dataset for intronic Flt1 SNPs, the focus of the present study was the exonic SNPs. As mentioned previously, neither of the two SNPs affects the predicted amino acid sequence, but our hypothesis states that they will have an effect on Flt1 exon 13-14 splicing efficiency. As an initial predictor of the effect of the Flt1 exonic SNPs, the web-based ESEFinder program, (release 3.0) was used [77]. This program predicts the efficiency of binding of four SR Proteins (SF2/ASF, SC35, SRp40, SRp55) to a potential ESE binding site in any given sequence [77]. Short (6-8 nucleotides), degenerate and partially overlapping consensus motifs for each of these proteins is considered (Figure 2.1) [77]. Output in the program is a set of "scores", with a higher score indicating a stronger match to empirically-determined SR binding motifs and thus a higher potential for binding of SR protein to each sequence [77]. A score is considered relevant if it is above a threshold value, which is specific to each SR protein (Figure 2.1) [77]. A change in the sequence (such as a SNP) can result in

differences in score values of ESE sites, with a higher difference score suggesting a more drastic effect caused by the SNP [85].

According to current SNP databases, the SNP found in Exon 13 (rs17537350) was not found in the Chinese population, but has a 3% presence of the minor allele (non-WT) in both Caucasian and Japanese populations (with no data for the African population). The SNP in Exon 14 (rs11843776), however, shows a 2% presence of the minor allele (non-WT) in the African population but no presence in the Caucasian, Chinese or Japanese populations. Both exonic SNPs result in no change in the amino acid sequence, with both SNP variants in exon 13 coding for a Threonine and both SNP variants in exon 14 coding for a Valine.

To test for potential effects of splicing, each SNP variant was run through the ESEFinder (for graphical illustration of results, see Figures 2.2 and 2.3). In Exon 13, the WT allele (G) predicts binding of SF2/ASF protein at different motifs surrounding the base, with varying strength. However, when the non WT allele is present (T), two SF2/ASF motifs disappear, with a third showing a slight increase in score. In addition, presence of the non WT allele introduces two new motifs for the SRp40 protein, thereby predicting a new set of predicted ESEs in the region. In Exon 14, the WT allele (C) predicts binding of SC35 and SRp40 proteins at different motifs with varying scores. However, when the non WT Allele is present (T), the predicted strength of binding changes for two SC35 motifs, thereby predicting a slightly lower strength of binding for predicted ESEs in the region.

Based on the potential effects predicted by the ESE finder program, we hypothesized that these SNPs would have an effect on the mRNA processing of the Flt1 gene. In order to test the hypothesis, three clones were constructed with the non wild type (WT) SNP, one with the Exon 13 SNP, one with the Exon 14 SNP and one with both SNPs. In addition, clones were designed in which the ESEs in the surrounding area of the SNPs were disrupted by mutations that likewise did not disrupt the amino acid sequence. These changes were done by changing bases and checking for ESEs that were silenced/created, until most ESEs were silenced. Thus, an additional three clones were created: one with mutations that resulted in most of the Exon 13 predicted ESEs silenced, one with mutations that resulted in most of the Exon 14 predicted ESEs

silenced, and one with most of the predicted ESEs silenced in both Exons 13 and 14. Details of bases changed for each mutation are found in Figure 2.4, while effects of mutation on ESEs are found in Figure 2.5 (Exon 13) and 2.6 (Exon 14)

Construction of WT Human Flt1 Minigene

In order to test the hypothesis, a human Flt1 minigene construct was assembled. The first attempt, utilizing the InFusion PCR Cloning Kit (Clontech, Catalog# 639606) to capture a 9185 bp PCR-amplified fragment spanning the Flt1 region of interest, was unsuccessful, likely owing to the large size of the insert to be cloned. A second attempt (also unsuccessful) involved amplifying the region from human genomic DNA with custom primers designed with built-in restriction endonuclease sites for *AgeI* (New England Biolabs (NEB)), and cloning into a linearized pcDNA expression vector cut with the compatible *BspEI* enzyme (NEB). An attempt to optimize the PCR with the Failsafe PCR Kit (EPICENTRE, Catalog # FS99060) indicated that the cloning step, rather than the PCR efficiency, was a limiting factor. This failure was also attributed to the large size of the Flt1 insert. The third and final unsuccessful attempt involved using a PCR-amplified gene fragment and use of the TA cloning kit (Invitrogen, Catalog # K4750-20). While this was partially successful in that an insert was cloned into the vector, the insert was truncated and showed several base pair anomalies at sequencing, which were attributed to PCR artifacts. This led to the decision to use unamplified sequence for the Wild Type model.

A human BAC clone (BAC RP11-502P18, BAC/PAC Resources, Children's Hospital, Oakland, CA) encompassing the Flt1 region of interest was obtained (Table 2.2). This clone contains the reference sequence that is considered 'Wild Type' (WT) for purposes of these comparative studies. The Flt1 region of interest (9185 bp) was excised from the 164,489 bp BAC insert by digestion with *EcoRV* (NEB) and purification of the Flt1 insert (Appendix A). The pLITMUS38 (pL38) holding vector (Figure 2.7) was linearized with *EcoRV* and the ends dephosphorylated with Calf Intestinal Phosphatase (CIP) enzyme (NEB, Appendix A). The insert was then ligated into the linearized pL38 clone and the constructed plasmid was purified. This purified plasmid was verified with restriction endonuclease mapping, as well as end-sequencing; all results were

consistent with the successful construction of pL38 vector with Flt1 insert (11.99kb in total).

Once the Flt1 insert was captured in the pL38 holding vector, it was then subcloned (by transforming with FusionBlue competent cells) into the expression vector pcDNA Intron A vector with CMV promoter (referred to as pcDNA in this document) [2] (Figure 2.8). The Flt1 insert was excised by cutting with two different restriction endonucleases, generating one blunt end (via *EcoRV*, NEB) and one overhanging end (via *NgoMIV*, NEB), with a loss of a 204 bp fragment from the Intron 12 portion (leaving a final length of 1763 bp for Intron 12 in the Flt1 fragment). The pcDNA vector was linearized with two compatible restriction endonucleases (*AleI* and *EcoRV*, NEB), with a loss of a 147 bp fragment from Intron A. Due to the nature of the excised ends, the insert could only be cloned in one direction, thus enabling the acquisition of the desired vector in one step. Structure was confirmed by initial restriction endonuclease digestion experiments and by complete bidirectional sequencing of the insert and vector in the final construct (total length: 15.225kb, Figure 2.8, detailed in Figure 2.9).

The expression of mRNAs encoded by the minigene plasmid was tested on cDNA amplified with primers annealing to the vector exons, CMV1 and CMV2 (spectrum of possible products summarized in Table 2.3). The CMV1-CMV2 product was the only one detected via end-point PCR amplification, which suggests that the minigene is correctly processing transcripts, albeit at a lower efficiency than expected.

Preparation of Mutant Minigenes

Once the construct was assembled, the mutants were constructed via site directed mutagenesis, whereby primers were used to amplify the region around the SNP for alteration of DNA sequence. Four different primers were used for each exon: two for SNP creation and two for ESE silencing (sequence of primers can be found in Table 2.4 and location on construct is shown in Figure 2.10). Due to the large size of the WT human Flt1 minigene clone, the exons and some flanking sequence were subcloned individually into smaller holding vectors based on pL38 (one for exon 13, one for exon 14) in order to facilitate the mutagenesis process.

The Exon 13 pL38 plasmid was created by cutting the Flt1 minigene with the *MfeI* restriction enzyme (NEB), thus excising a 4934 bp fragment containing Flt1 exon 13. The pL38 holding vector was cut with the same enzyme, its ends were dephosporylated with the CIP enzyme and then the Flt1 fragment was ligated into this prepared vector (Figure 2.11). The new exon 13 plasmid was verified with restriction enzyme digestions and sequencing of the ends, to confirm presence of entire fragment as well as orientation.

The Exon 14 pL38 holding vector was created by cutting the Flt1 minigene with the *EcoRV* and *NheI* restriction enzymes (NEB), thus excising a 2762 bp fragment containing Flt1 exon 14. The pL38 holding vector was cut with the same enzymes (without phosphorylation this time, as it was cut with two non-compatible enzymes) and then the Flt1 fragment was ligated into this prepared vector (Figure 2.12). The new exon 14 plasmid was verified with restriction enzyme digests and sequencing of the ends, to confirm presence of entire fragment. Orientation in this case was not a concern, as the insert could only be ligated in the correct orientation due to the different enzymes used.

Once the holding vectors with each exon were prepared, the Flt1 sequence was mutated via site-directed mutagenesis. The mutated sequences were amplified from each holding vector with the primers required (Table 2.4) in the thermocycler in two successive PCR reactions (details of thermocycler conditions in Appendix A). The primary PCR reaction produced two fragments that overlapped at the site of the mutation, while the secondary PCR reaction extended each end of these fragments to produce a long double stranded mutated DNA fragment [106]. This mutated DNA fragment was then replaced in each of the holding vectors to replace the wild type fragment (Figures 2.13 and 2.14 for exons 13 and 14, respectively).

Once the mutated DNA was cloned into fragments, the clone was characterized with restriction endonuclease mapping, and the entire PCR amplified portion was sequenced not only to check for presence of mutations, but also to ensure no other base pair changes were made by way of PCR artifacts. If any mutations were found besides the ones designed in the experiment, the primary and secondary PCR steps were repeated, and cloned all over again until a clone was produced with no base pair changes except intended mutations.

The next step entailed transferring these mutated fragments back into the pcDNAIntA expression vector by replacing the WT sequence in the Flt1 minigene with the mutated one prepared in the respective holding vector for each exon. This was done in the same manner the pL38 holding vectors were produced, but in reverse (as explained in Figures 2.10 and 2.11). Exon 13 mutated regions were excised out of each of the pL38 mutant clones (SNP and ESE) back into pcDNA by cutting exon 13 mutant pL38 plasmids and full length Flt1 pcDNA with *MfeI*, and ligating the excised mutant fragment into the *MfeI* linearized vector, thereby replacing the WT sequence. Exon 14 mutated regions were excised out of each of the pL38 mutant clones (SNP and ESE) back into pcDNA by cutting both exon 14 mutant pL38 plasmids and full length Flt1 pcDNA with *EcoRV* and *NheI* and ligating the excised mutant fragment into the *EcoRV* and *NheI* linearized vector, thereby replacing the WT sequence.

However, this cloning step proved to be a far more problematic task than anticipated, mainly due to the large size of the clone (about 15Kbp). Many unsuccessful attempts were made to optimize this step, including varying concentrations of insert and vector used, changing endonucleases, changing ligation conditions, changing ratios of insert:vector, changing amount of transformed ligated sample, changing type of competent cells used (chemical and electrocompetent) and even the use of specialized kits (such as Stratagene's QUIK XL II, Catalog #200516 and QUIKLightning Kits, Catalog #21058).

Ultimately, a combination of optimized steps was used that eventually resulted in successful cloning: ensuring more complete restriction digestion for both "insert" and "vector," modifying the gel extraction process to maximize yield, eluting with 1X TE buffer rather than nuclease-free water, and the use of FusionBlue competent cells (Clontech) for transformation (Appendix A).

For exon 13, an additional step was employed to increase the efficiency cloning. Because the vector was cut with only one enzyme, the cohesive ends tended to ligate back together even when dephosphorylated. To create an easily identifiable 'background' vector, the *MfeI* fragment was excised from the full length Flt1 minigene, and re-ligated (Figure 2.15). This smaller plasmid was purified, linearized and then used to receive the excised mutated exon 13 fragments from the pL38 holding vectors. The

new full-length mutants could then be readily distinguished by PCR from the desired clones containing insert. Such a step was not required (nor available) for the exon 14 mutants, as the use of two enzymes to excise the exon 14 fragment prevented such a high background.

Making the double mutants was a more straightforward task. The two pcDNAIntA exon 14 mutants (SNP and ESE, respectively) were used as the “insert”, with a large fragment containing exon 14 region excised using *BspEI* and *EcoRV*. The two pcDNAIntA exon 13 mutants (SNP and ESE) were used as the “vector,” removing the wild type exon 14 fragment to receive the mutated exon 14 fragment, thereby producing a plasmid with mutations in both exons (Figure 2.16).

Each of the clones was characterized by a restriction enzyme digest (Figure 2.17) and sequencing (Figures 2.18 and 2.19, for exons 13 and 14, respectively). Once all clones had been prepared, they were all re-grown from bacterial stocks previously frozen (Table 2.5 lists lab nomenclature of clones), then purified using the Zyppy Endofree plasmid prep kit (Appendix A) and stored at 4°C until ready to use for transfection. Restriction mapping and sequence confirmation was performed on all plasmids prior to the initiation of transfection experiments.

Flt1 and sFlt1 Expression in Transient Transfections

The assembled Flt1 minigene constructs were transfected (Appendix A) into mouse embryonic fibroblast NIH/3T3 cells in order to test effects of the SNP and ESE mutations on Flt1 and sFlt1 mRNA expression. Transfection was optimized, and several test transfections were run to ensure the system was working as anticipated. This included a test of transfection efficiency by concurrent transfection of the Flt1 minigene along with a parallel transfection of Green Fluorescent Protein (GFP) in order to visualize the transfected cells.

Once it was determined that the conditions, primers and reagents were satisfactory for transfection, all 6 mutants were transfected alongside the WT clone and an empty pcDNAIntA vector (i.e. with no Flt1 insert) as a control. This was repeated four times, each set transfected on a separate day and from a separate flask and/or passage number, in order to test reproducibility of results. All transfectants were harvested after

48 hours, and RNA isolation immediately performed. This was followed by DNaseI digestion, purification, and quantification of RNA and cDNA synthesis. In addition to cDNA, non-reverse transcribed (No RT) controls were also prepared, and both cDNA and its respective No RT control were then analyzed via QPCR to detect levels of Flt1 and sFlt1 mRNA products. Conventional PCR was also performed for each sample to confirm presence of Flt1 and sFlt1 fragments (Please refer to Appendix A for details of each procedure).

Results

A representative qPCR reaction profile for each of Flt1 and sFlt1, as well as dissociation curves for each target, can be found in Figures 2.20 and 2.21, respectively. The results of the transfections are summarized in Figure 2.22 (Change in ratio of Flt1 to sFlt1 levels), Figure 2.23 (Change in Flt1 levels normalized to Neomycin) and Figure 2.24 (Change in sFlt1 levels normalized to Neomycin).

SNPs mutants decrease Flt1:sFlt1 ratio, due to a decrease in Flt1 levels

As is evident in Figure 2.22 (Change in ratio of Flt1 to sFlt1 levels), the ratio appears to decrease with the presence of one or both SNP mutations. There is a decrease from WT expression of 40% and 46% of WT expression with a SNP in exon 13 or 14, respectively (standard error of mean (SEM) is 0.11 and 0.26, respectively), and a decrease of 41% from WT expression with a SNP in both exons (SEM = 0.07).

This difference in ratio in the presence of the exonic SNPs is mainly due to a decrease in Flt1 expression levels. In the exon 13 SNP clone, the decrease in Flt1:sFlt1 is due to a decrease of Flt1 expression (Figure 2.23) to 30% of wild type expression (SEM = 0.12), as sFlt1 expression does not seem to be affected by this SNP (Figure 2.19). In the presence of exon 14 SNP, both Flt1 and sFlt1 expression are affected: Flt1 levels decrease to 32% of wild type expression (SEM = 0.15) and, and sFlt1 increases by about 42% when compared to wild type (SEM = 0.6). When the SNP mutation occurs in both exons, the decrease in the ratio of Flt1:sFlt1 observed is due to a 46% decrease

from WT expression in Flt1 (SEM = 0.07%), as the SNP in both exons does not appear to affect sFlt1 expression.

ESE Mutants have no consistent effect on Flt1 and sFlt1 levels

Unlike the SNP mutants, the presence of the mutations in the predicted ESEs in surrounding exons do not have a consistent effect on Flt1 and sFlt1 levels when present in either exon, as compared to WT expression levels (Figure 2.24). However, an ESE mutation in both exons increases the Flt1:sFlt1 ratios by slightly over 100% more than WT expression (SEM = 0.3). This change in ratio is not due to Flt1 levels, as they don't appear to change when ESEs in both exons are mutated, but there is a decrease of sFlt1 levels to 58% of WT expression (SEM = 0.3), which would account for the observed change in the Flt1:sFlt1 ratio.

Statistical Analysis

The WT was compared to the three SNP mutant clones and the double ESE mutant clone using ANOVA followed by Dunnett Procedure for Multiple Comparisons. Statistical significance was set to alpha = 0.05, and all analysis were performed using SAS. (Please see Appendix B for output of Statistical results).

Overall, the mutants had a statistically significant effect on the Ftl1:sFlt1 ratio (p value: 0.0048). Compared to the WT, the single SNP mutants appear to have a some effect on Flt1:sFlt1 levels (nominal p-values before correction for multiple comparisons was 0.0436 and 0.0253 for Exon 13 SNP and Exon 14 SNP, respectively). The nominal p-value for the double SNP mutant was borderline (0.0544), and clearly not significant for the double ESE mutant (0.1015). When adjusted for multiple comparisons, all statistical significance disappeared (p values: 0.1322, 0.0796, 0.16 27and 0.2826 for Exon 13, Exon 14, Double SNP and Double ESE mutants, respectively).

Similarly, the mutants had a statistically significant effect on Ftl1 levels (normalized to Neo) (p value: 0.0047). Compared to the WT, all SNP mutants had a statistically significant effect (p values were 0.0033, 0.0047 and 0.0365 for Exon 13 SNP, Exon 14 SNP and double SNP mutants, respectively). Conversely, the double ESE mutant does not appear to have a significant effect (p value = 0.7972). When the p

value were adjusted for multiple comparisons, significance of the change in levels due to the single SNP mutants remained apparent (p values were 0.0113 and 0.0158 for Exon 13 SNP and Exon 14 SNP, respectively) but not so for the double SNP mutant (p value = 0.1121).

The overall effect of the mutants on sFtl1 levels (normalized to Neo) was not statistically significant (p value: 0.6756). Individually, none of the clones showed statistical significance before (p values: 0.3154, 0.5195, 0.8891 and 0.2350 for Exon 13 SNP, Exon 14 SNP, double SNP and double ESE mutants, respectively) or after adjustment for multiple comparisons (p values: 0.6947, 0.9097, 0.9997 and 0.5635 for Exon 13 SNP, Exon 14 SNP, double SNP and double ESE mutants, respectively).

Discussion

The occurrence of discrete genetic differences among individuals is a fundamental contributor to onset and progression of a number of human diseases. The simplest form of genetic difference, SNPs, potentially can affect any aspect of regulated gene expression: rates of transcription, RNA processing, stability and translation, and protein function. The present study was undertaken to test the hypothesis that exonic SNPs in the Flt1 gene can alter relative rates of RNA processing leading to mRNAs encoding full-length Flt1 or its truncated variant sFlt1. It was anticipated that if the SNPs were in regulatory regions for cleavage-polyadenylation or splicing mechanisms, the mutations would affect the ratios of sFlt1 and Flt1 mRNAs, respectively. Based on the ~40% reduction in Flt1:sFlt1 mRNA ratios observed with the presence of a non WT allele, it appears the hypothesis is supported and that the SNPs tested, in the context of our minigene experimental system, do indeed lie in regions that help determine relative rates of RNA processing for these transcripts.

Moreover, the SNPs appear to decrease the exon 13/14 splicing activity of the Flt1 minigene, as the presence of SNPs results in a decrease in Flt1 mRNA relative to the internal Neo control. Consistent with previous studies in our lab involving mutated

splicing determinants [107], there was not uniformly detectable reciprocity between levels of Flt1 and sFlt1 mRNAs. That is, the decreases in Flt1:Neo were not accompanied by increases in sFlt1:Neo for the mutations made.

While mutating the exon 13 or exon 14 area around each respective SNP appeared to have little effect on the Flt1 surrogate mRNA levels, and therefore little effect on splicing, it is important to note that the ESE mutation does not introduce the non WT allele in either case (but rather keeps the wild type allele at that base), and therefore is consistent with the importance of the SNP itself. The lack of a consistent effect of ESE mutations in these clones could be due to several reasons. For one, the mutations did not completely silence all predicted ESEs in the region (as evident in Figures 2.5 and 2.6), and it is possible that the functional ESEs were the ones that failed to be silenced. Another possibility is that all predicted ESEs in these regions are simply nonfunctional; further testing (by silencing the rest of the ESE mutations in the region, which was not possible to do manually) would be needed to confirm this. Furthermore, there is very little information available for exonic splice silencers (ESS), which may very well have a more dominant effect than the predicted ESEs in the region. These exonic splice silencers may be present naturally, and been inadvertently silenced through the mutations made to silence ESEs: silenced ESEs and silenced ESS would therefore cancel the effects of each other, producing no effects.

While such a minigene approach is very valuable in validating the effects of changes in the DNA sequence, including single base pair changes, there are several limitations. In this case, the entire Flt1 is far too large to be studied in such a manner easily, so the exon13-intron13-exon14 region within our pcDNAIntA plasmid may behave in a manner that differs from its behavior in its native context. Even if the gene fragment behaves similarly to its native context, base pair changes made may affect the minigene system in a manner that differs from how these same changes would affect the full length gene in it's native context [108]. Another caveat is that splicing efficiency is linked to the promoter, so a change that affects splicing using the plasmid's promoter may not do so when using the native promoter [108].

A major limitation in this study is the assumption that our minigene experimental system, with the Flt1 and sFlt1 surrogates, is mimicking the true expression levels of

Flt1 and sFlt1 levels in endothelial cells. However, the minigene model is a very valuable tool to demonstrate whether or not there is a potential effect, and further study in endothelial cell models can then be pursued. This would be attempted by purifying endothelial cells from an adipose tissue biopsy taken from an individual with a non WT SNP variant, as well as from an individual with a WT SNP variant, and commencing with analyzing expressing levels of these purified endothelial cells to compare expression levels in each case.

The endogenous Flt1:sFlt1 delta CT was detected to be -2.04 in HUVEC cells, which is lower than the Flt1:sFlt1 ratio of the WT minigene (an average of 1.87 for the four transfection replicates). This difference suggests a lower splicing efficiency in the minigene transfectants than that of the endogenous control.

In conclusion, the exonic SNPs result in a decrease of Flt1:sFlt1 levels in our minigene model (mainly due to a decrease is Flt1 levels). This indicates that SNPs do lie in important ESE sites, and by their presence, appear to decrease the strength of the ESE binding sites, causing the splicing machinery to fail to recognize exon 14 as efficiently and thereby, decreasing Flt1 levels. In essence, the SNPs in the exons do appear to have functional effects, and it is thus valid to assume that SNPs in the intron 13 region (a region already determined to be important in Flt1 and sFlt1 processing) would also have a functional effect. Therefore, a thorough study of the functional effects of SNPs in the area of the exon13-intron-13-exon14 region is an extremely important area of study, as we seek to understand the factors determining Flt1 and sFlt1 expression.

Table 2.1: Screen of SNPs in Flt1 region in initial stages of study (2006). SNPs in exons are highlighted. Neither exonic nor intronic SNPs are validated, and most SNPs have no population data associated. †Listed as Double Hit in SNP Browser but not NCBI, *Also listed as part of a 2bp insertion (GA/TT, rs35409953 with no frequency data), Val = Validation status (Nt= Not Validated, D = Double hit, V = Validated), Source: N = NCBI, S = SNP Browser, B = Both

SNP ID	SNP	WT	Val	Source	African		Caucasian		Chinese		Japanese	
rs17086623	C/T	C	D	N	1.00	0.00	1.00	0.00	0.92	0.08	N/A	
rs34483960	-/A	-	Nt	N	No Data							
rs17086620	A/G	A	D	B	N/A		0.93	0.07	0.99	0.01	1.00	0.00
rs7983774	A/G	G	D	N	0.16	0.84	0.28	0.72	0.35	0.65	0.42	0.58
rs17537350	C/T	C	D	B	N/A		0.97	0.03	1.00	0.00	0.97	0.03
rs2296283	A/G	G	D	B	N/A		0.58	0.42	0.63	0.37	0.52	0.48
rs2296284	A/G	G	D	B	0.14	0.86	0.27	0.73	N/A		N/A	
rs5802479	-/A	A	Nt	N	No Data							
rs17086618	A/G	A	Nt	B	0.98	0.02	1.00	0.00	1.00	0.00	1.00	0.00
rs4771246	C/T	C	D	B	1.00	0.00	0.95	0.05	1.00	0.00	0.97	0.03
rs9551465	A/T	A	D	B	0.93	0.07	0.41	0.59	0.37	0.63	0.48	0.52
rs17086617	C/T	T	D	B	0.15	0.85	0.28	0.72	0.35	0.65	0.44	0.56
rs7337610	C/T	T	D	B	0.18	0.82	0.35	0.65	0.36	0.64	0.48	0.52
rs11400597	-/A	-	Nt	N	No Data							
rs7325639	C/T	C	D	B	No Data							
rs17626077	G/T	G	D	B	0.92	0.08	0.94	0.06	1.00	0.00	1.00	0.00
rs35630136	-/G	-	Nt	N	No Data							
rs3088111	C/T	T	D	B	0.08	0.92	0.00	1.00	0.00	1.00	0.00	1.00
rs3751397	A/T	A	D	B	0.12	0.88	0.58	0.42	0.58	0.42	0.53	0.47
rs3751396	A/G	A	D	B	1.00	0.00	1.00	0.00	0.99	0.01	0.97	0.03
rs7319123	A/C	A	D	B	0.18	0.82	0.00	1.00	0.00	1.00	0.00	1.00
rs3794404	A/C	A	D	B	1.00	0.00	1.00	0.00	0.99	0.01	0.98	0.02
rs11843776	A/G	G	D	B	0.02	0.98	0.00	1.00	0.00	1.00	0.00	1.00
rs3751395	A/C	C	D	B	N/A		0.39	0.61	0.61	0.39	0.64	0.36
rs34113415	-/T	-	Nt	N	No Data							
rs11347871	-/A	A	Nt	N	No Data							
rs11294603	-/A	A	Nt	N	No Data							
rs12871995	A/C	C	Nt	N	No Data							
rs12872498	A/C	C	Nt	N	No Data							
rs7140067	C/T	C	D	B	0.91	0.09	0.97	0.03	1.00	0.00	1.00	0.00
rs13378470	C/T	C	Nt	N	No Data							
rs6491281	C/T	T	Nt	N	No Data							
rs13378467	C/T	C	Nt	N	No Data							
rs6491280	A/G	G	D†	B	0.99	0.01	1.00	0.00	1.00	0.00	1.00	0.00
rs9551464*	A/T	A	Nt	N	No Data							
rs9551463*	G/T	G	Nt	N	No Data							

Figure 2.1 [77]: Consensus motifs for SR protein considered in the ESEfinder web based program. As an initial predictor of the effect of the Flt1 exonic SNPs, the web-based ESEFinder program, (release 3.0) was used. This program predicts the efficiency of binding of four SR Proteins (SF2/ASF, SC35, SRp40, SRp55) to a potential ESE binding site in any given sequence. The illustration depicts the short (6-8 nucleotides), degenerate and partially overlapping consensus motifs for each of these proteins.

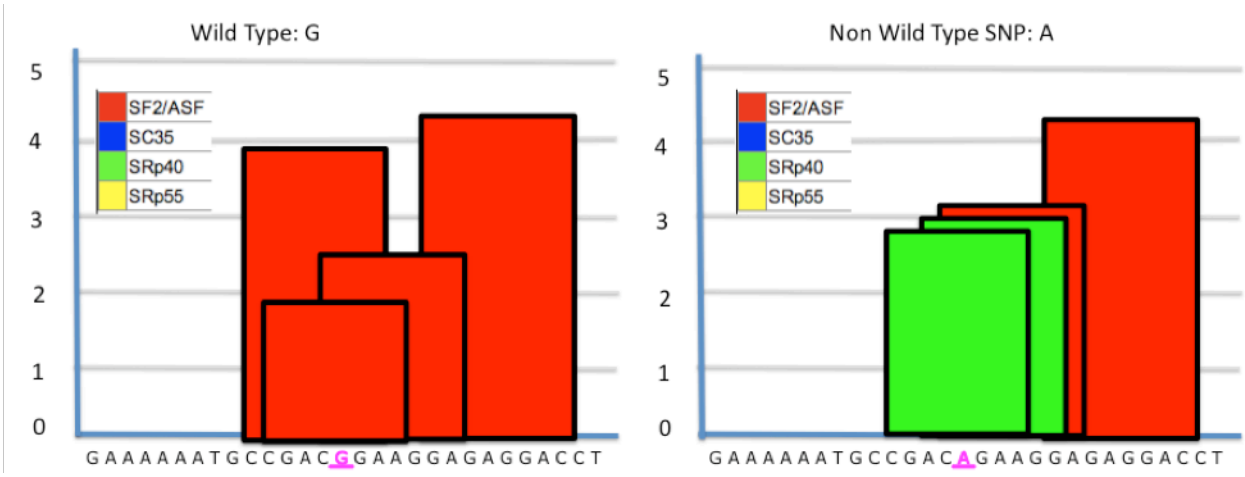


Figure 2.2: Effect of SNP in Exon 13 on predicted ESEs in the region. Illustration on the left shows WT allele, while right shows non WT allele (base in question is highlighted in pink font). The WT allele (G) shows predicted binding of SF2/ASF protein at different motifs with varying scores. However, when the non WT allele is present (T), two SF2/ASF motifs disappear, with a third showing a slight increase in score. In addition, presence of the non WT allele introduces two new motifs for the SRp40 protein, thereby predicting a new set of predicted ESEs in the region. (The proteins are indicated as blocks, scattered on the X axis (the DNA sequences). The width of the bars indicates the length of the motif for each protein, and the height of the bar indicates the strength of binding at indicated motif. A higher score indicates a stronger potential for binding).

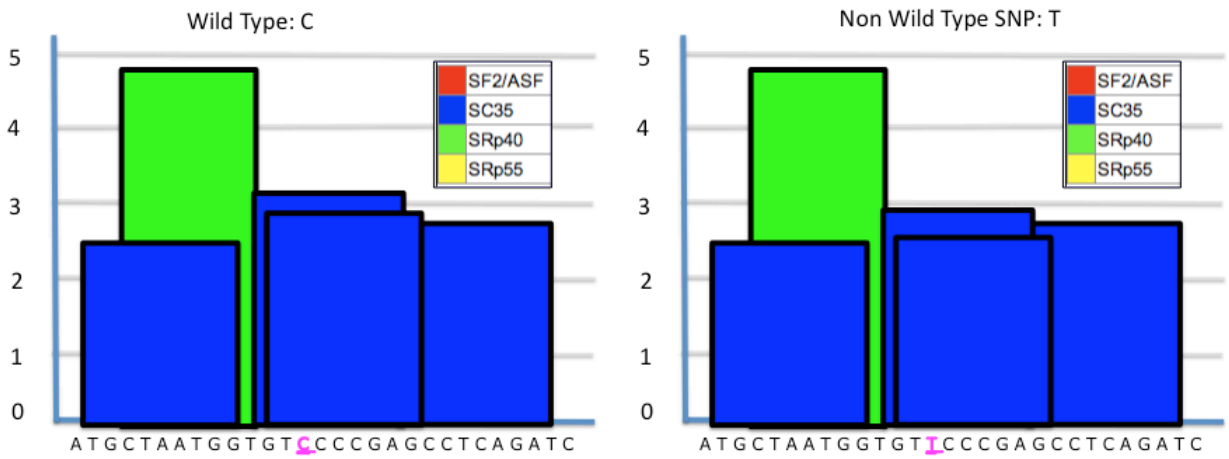


Figure 2.3: Effect of SNP in Exon 14 on predicted ESEs in the region. Illustration on the left shows WT SNP, while right shows non WT SNP (base in question is highlighted in pink font). The WT allele (C) predicts binding of SC35 and SRp40 proteins at different motifs with varying scores. However, when the non WT Allele is present (T), the predicted strength of binding changes for two SC35 motifs, thereby predicting a slightly lower strength for predicted ESEs in the region. (The proteins are indicated as blocks, scattered on the X axis (the DNA sequences). The width of the bars indicates the length of the motif for each protein, and the height of the bar indicates the strength of binding at indicated motif. A higher score indicates a stronger potential for binding).

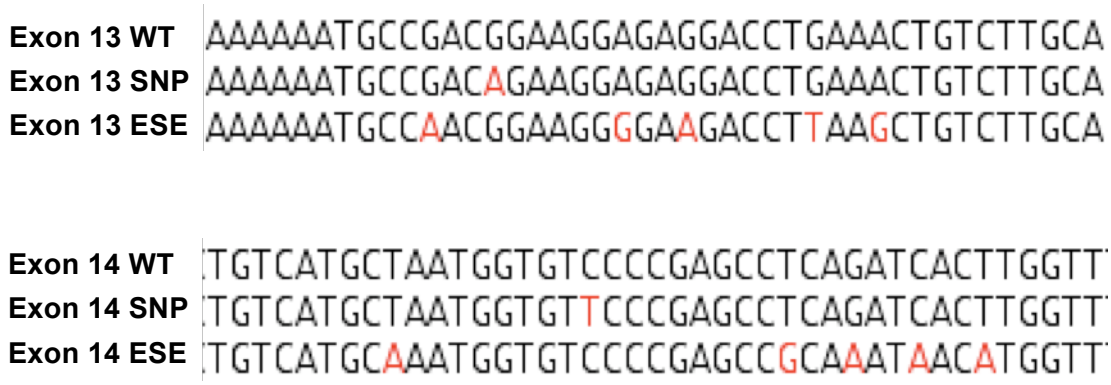


Figure 2.4: Bases changed in mutant clones. Top illustrations shows bases changed from WT sequence to create Exon 13 SNP and ESE mutants, respectively. Bottom illustration shows bases changed from WT sequence to create Exon 14 SNP and ESE mutants.

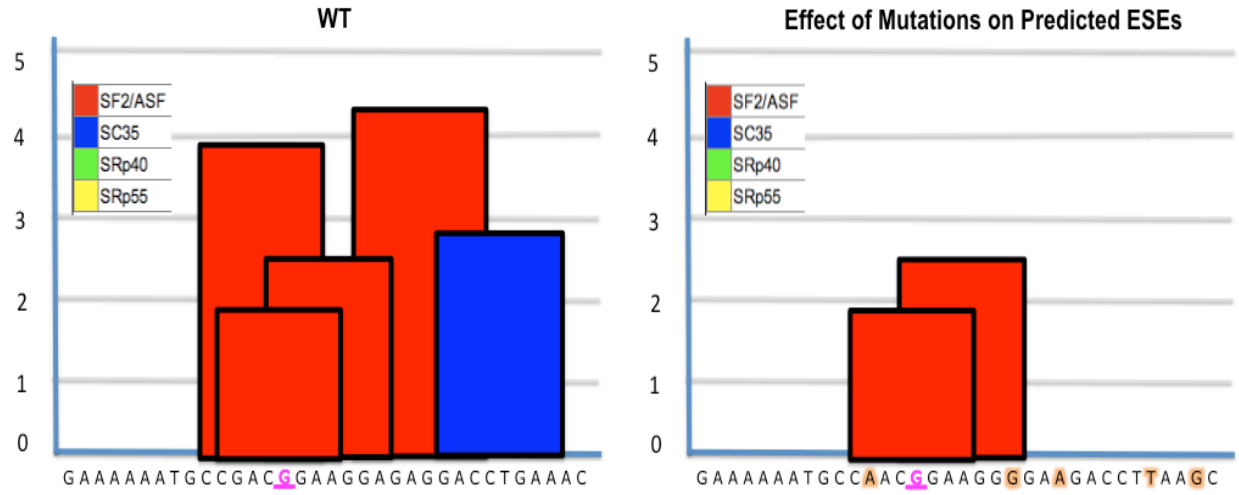


Figure 2.5: Effect of Mutations for silencing predicted ESEs in Exon 13 SNP region. Illustration on left shows predicted ESEs present in WT sequence, while illustration on right shows predicted ESEs after mutation. SNP location is indicated by pink base, while mutated bases to silence ESEs are indicated with orange highlight. As is evident, two predicted ESEs were not silenced with mutations. (The proteins are indicated as blocks, scattered on the X axis (the DNA sequences). The width of the bars indicates the length of the motif for each protein, and the height of the bar indicates the strength of binding at indicated motif. A higher score indicates a stronger potential for binding).

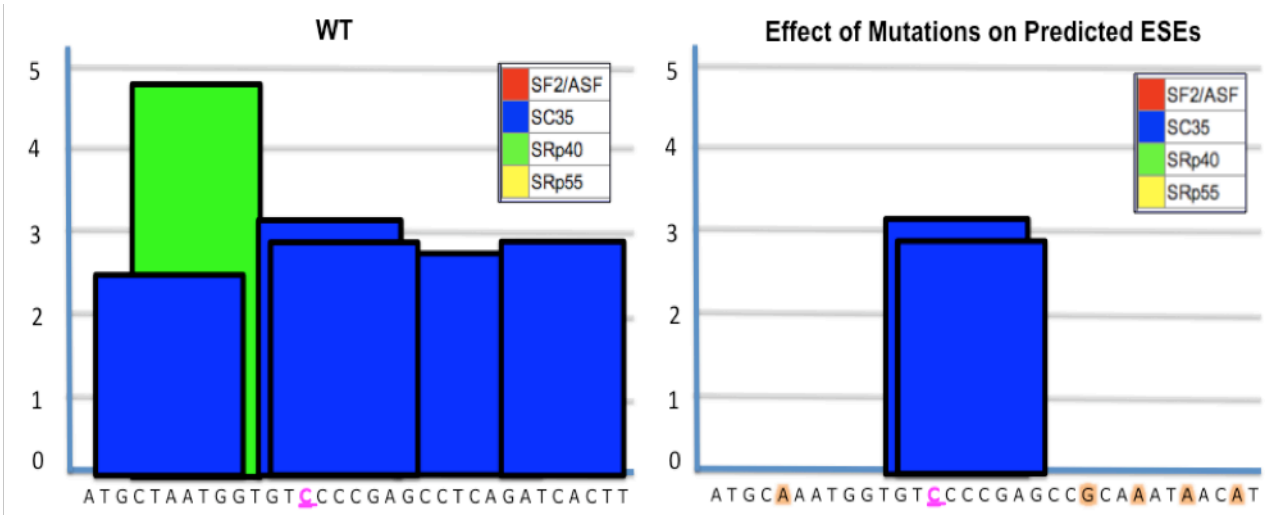


Figure 2.6: Effect of Mutations for silencing predicted ESEs in Exon 14 SNP region. Illustration on left shows predicted ESEs present in WT sequence, while illustration on right shows predicted ESEs after mutation. SNP location is indicated by pink base, while mutated bases to silence ESEs are indicated with orange highlight. As is evident, two predicted ESEs were not silenced with mutations. (The proteins are indicated as blocks, scattered on the X axis (the DNA sequences). The width of the bars indicates the length of the motif for each protein, and the height of the bar indicates the strength of binding at indicated motif. A higher score indicates a stronger potential for binding).

Table 2.2: Composition of Flt1 region excised from BAC Clone. The Flt1 region of interest excised was made up of the entire sequence of Exon 13-Intron13-Exon14, as well as portions of surrounding exons. This was first clones into the pL38 holding vector, and then further cloned in the pcDNAIntA expression vector, to make the Wild Type Human Flt1 Minigene Construct

Flt1 region	Length In Construct (bases)	Full Length in Gene (bases)
Intron 12	1967	6855
Exon 13	309	309
Intron 13	4764	4764
Exon 14	147	147
Intron 14	1992	27199

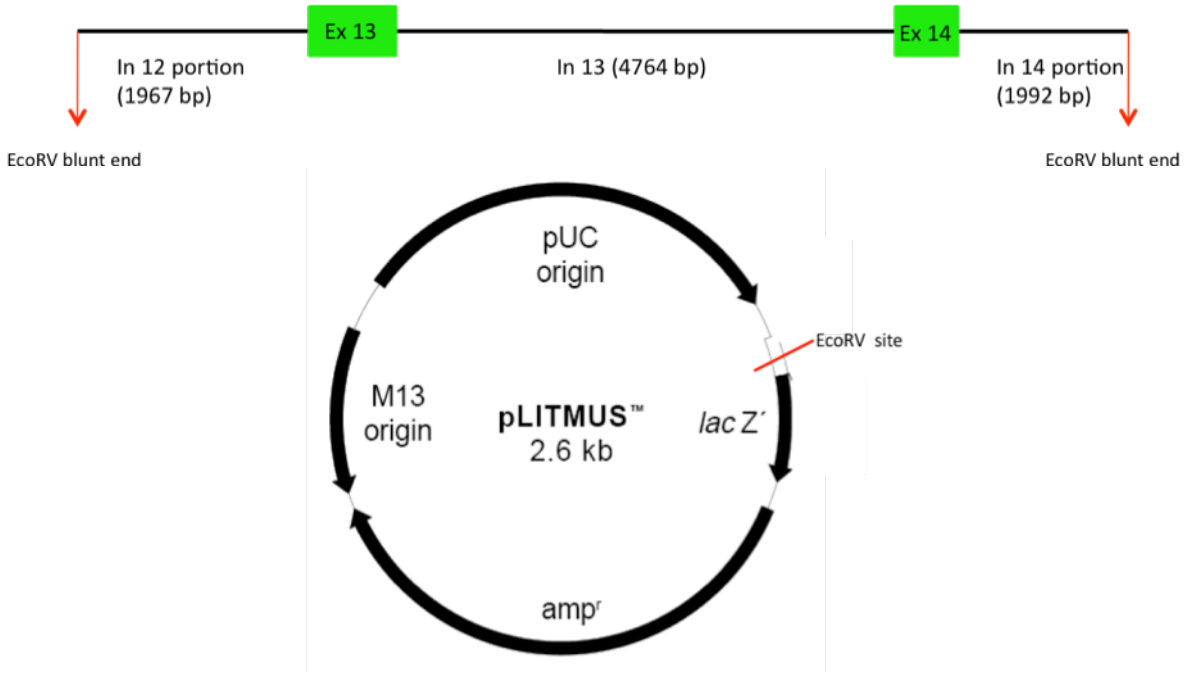


Figure 2.7: pLITMUS38 vector and Flt1 insert excised from BAC clone. The Flt1 region of interest (top) was excised from the BAC insert by digestion with *EcoRV* (NEB) and purifying the Flt1 insert. The pLITMUS38 holding vector was linearized with *EcoRV* and the ends dephosphorylated with Calf Intestinal Phosphatase (CIP) enzyme. The insert was then ligated into the linearized pL38 clone and the successfully constructed plasmid was purified (11.99kb in total) and characterized with restriction endonuclease digestion, as well as end-sequencing.

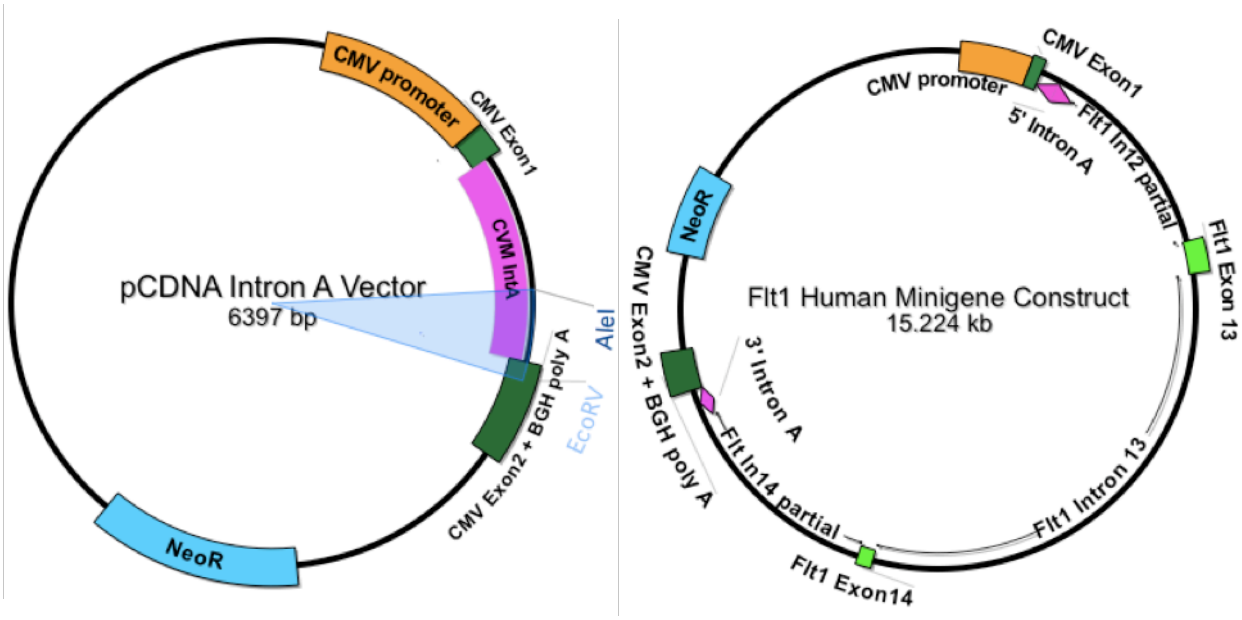


Figure 2.8: Cloning the Flt1 insert into the pcDNAIntA expression vector, from constructed pL38 clone. Left: Expression vector pcDNA Intron A, which was linearized with *EcoRV* and *AleI.* Excised portion of pcDNAIntA is highlighted, which is the site of insertion of the Flt1 fragment. Right: Assembled Flt Minigene Construct, with Flt1 insert cloned into pcDNAIntA vector. (NeoR = Neomycin resistance gene)

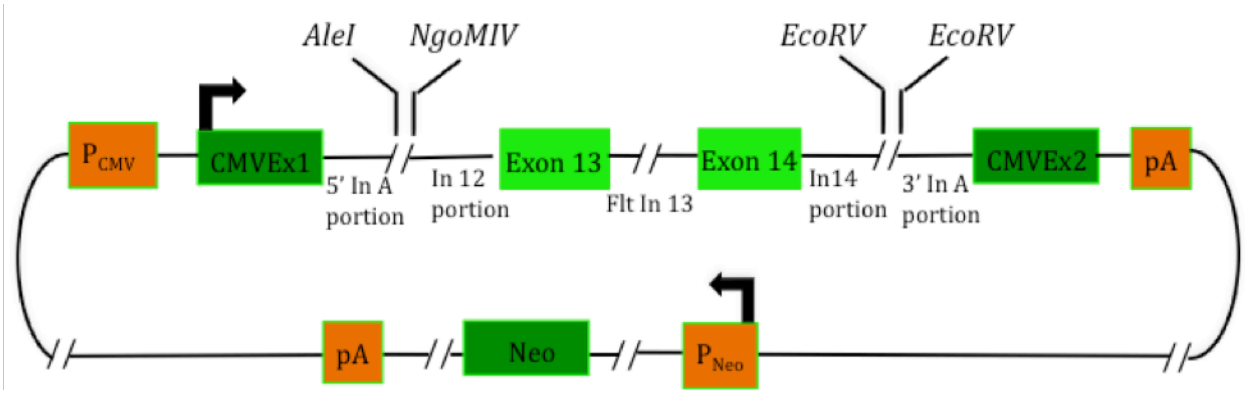


Figure 2.9: Details of assembled human Flt1 Minigene Construct. Figure not drawn to scale. P_{CMV} = CMV promoter, P_{Neo} = Neomycin promoter, CMVEx = CMV Exon in pcDNA, In= intron , pA = polyadenylation signal, Neo = Neomycin gene. Arrow denotes transcription start site

Figure 2.10: Location of primers used for site directed mutagenesis. Each set is used to create the mutation, while Primers 1 and 4 and Primers 5 and 8 are used to extend the double stranded mutant DNA. Details of primers are in Table 2.4

Table 2.3: Testing Expression of the Minigene. Primers in the vector exons, CMV1 and CMV2, were used to detect a spectrum of possible products

Predicted Product	Primers to Detect	PCR amplicon Size	Detected
CMV1, Ex13, Ex14, CMV2	BH 302 and BH 246	843 bp	No
CMV1, Ex13	BH 302 and BH 387	446 bp	No
CMV1, Ex14, CMV2	BH 302 and BH 246	555 bp	No
CMV1, Ex13, CMV2	BH 302 and BH 246	717 bp	No
CMV1, CMV2	BH 302 and BH 246	408 bp	Yes

Table 2.4: Details of primers used in site directed mutagenesis. Name, use of mutation and sequence of each primer is indicated, as well as location of each primer on Figure 8 is indicated

Primer Name	Location in Fig 2.8	Description of Primer	Sequence of Primer
BH 562	1	5' Exon 13 Nonmutant Primer	GAGGCCGAGGAGCATGGATCACCT
BH 585	2	3' Exon 13 SNP Mutant Primer	CAGGTCCTCTCCTTCtGTCGGCATTTTTTCC
BH 584	3	5' Exon 13 SNP Mutant Primer	GGAAAAAATGCCGACaGAAGGAGAGGACCTG
BH 542	4	3' Exon 13 Nonmutant Primer	CAGGAGAAAACAGCAAGAGCAACAA
BH 562	1	5' Exon 13 Nonmutant Primer	GAGGCCGAGGAGCATGGATCACCT
BH 589	2	3' Exon 13 ESE Mutant Primer	GCAAGACAGcTTaAGGTCtTCcCCtTCcGTtGGCATTTTTTCCAAG
BH 588	3	5' Exon 13 ESE Mutant Primer	GCCaACgGAaGGgGAaGACCTtAAgCTGTCTTGCACAGTTA
BH 542	4	3' Exon 13 Nonmutant Primer	CAGGAGAAAACAGCAAGAGCAACAA
BH 303	5	5' Exon 14 Nonmutant Primer	CAACAAGAGGGGAAAGGAATGACT
BH 587	6	3' Exon 14 SNP Mutant Primer	GATCTGAGGCTCGGGaACACCATTAGCATGAC
BH 586	7	5' Exon 14 SNP Mutant Primer	GTCATGCTAATGGTGTtCCCGAGCCTCAGATC
BH 570	8	3' Exon 14 Nonmutant Primer	GAGCCAAGATGGTGTCACTGTAC
BH 590	5	5' Exon 14 ESE Mutant Primer	GCaAATGGTGTcCCCGAgCCgCAaATaACaTGGTTTAAAAACAAC
BH 570	6	3' Exon ESE Nonmutant Primer	GAGCCAAGATGGTGTCACTGTAC
BH 591	7	5' Exon 14 ESE Mutant Primer	AAACCAtGTtATtTGcGGcTCGGGgACACCATTtGCATGACAGTCTAAAG
BH 303	8	3' Exon 14 Nonmutant Primer	CAACAAGAGGGGAAAGGAATGACT

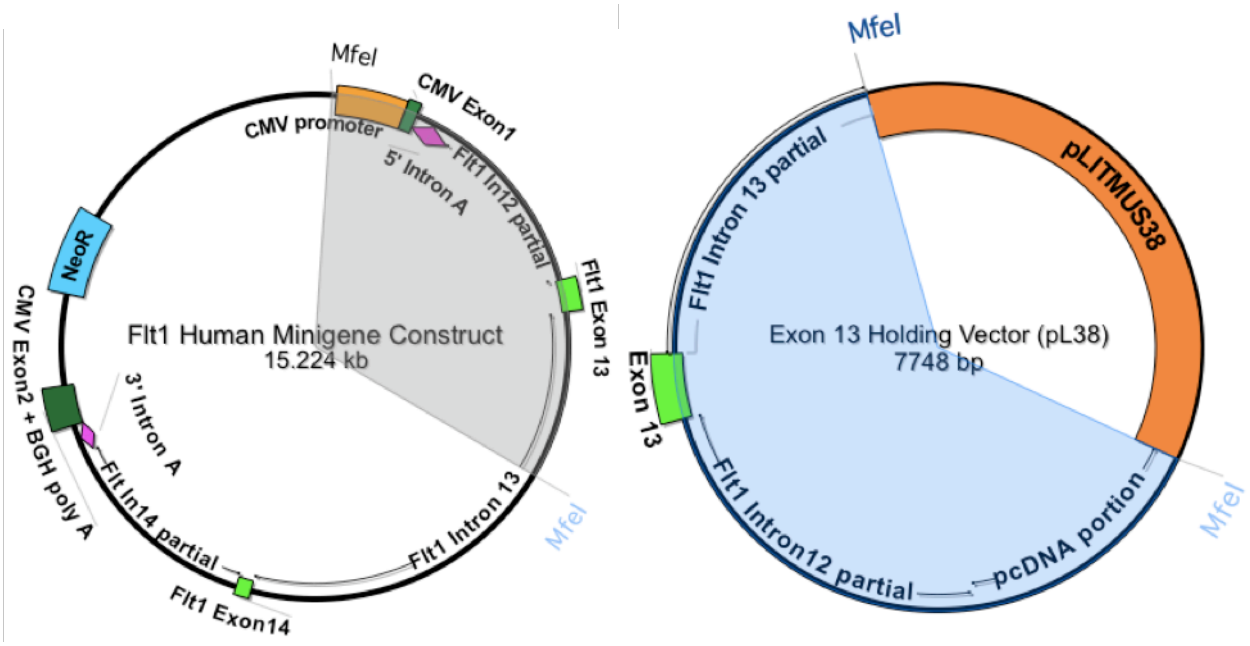


Figure 2.11: Creating Exon13 pL38 vector. In order to facilitate the exon 13 mutagenesis process, the surrounding region was subcloned into a pLITMUST38 holding vector, by excising fragment containing exon 13 (4934 bp) from the full length Flt1 pcDNA minigene (left) with *MfeI*, and ligating into plL38 linearized with *MfeI*. The new exon 13 pL38 plasmid (right) was verified with restriction enzyme digests and sequencing of the ends, to confirm presence of entire fragment as well as orientation. Cloned portion is highlighted in both donor and acceptor vector.

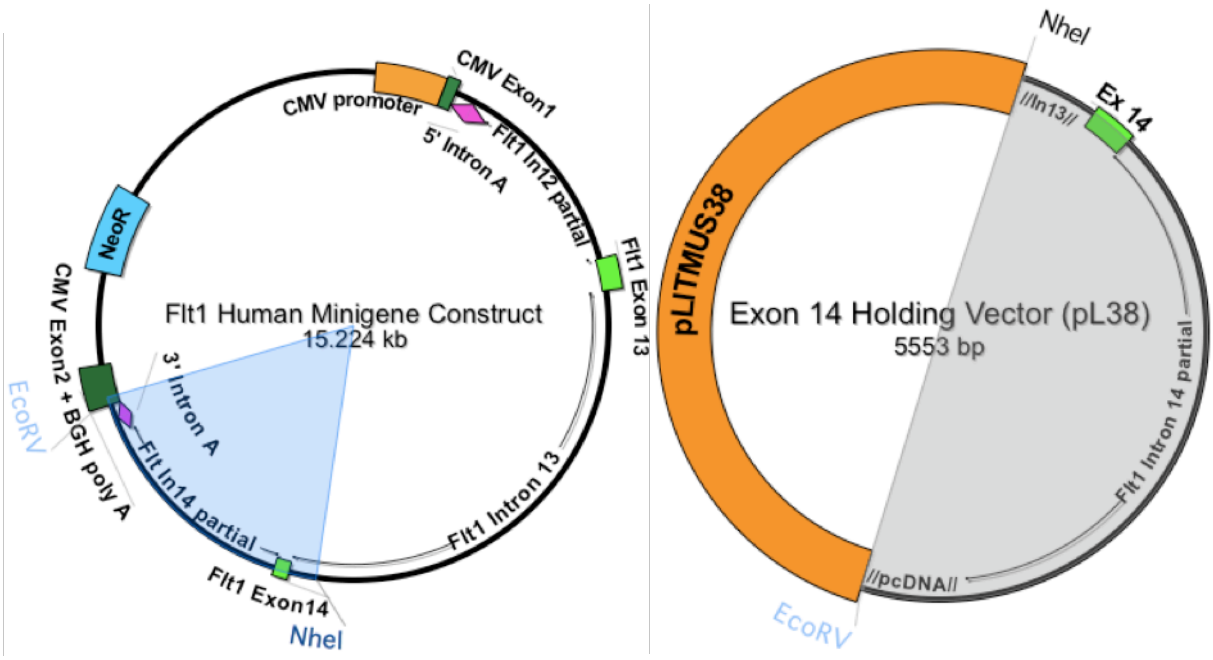


Figure 2.12: Creating Exon14 pL38 vector. In order to facilitate the exon 14 mutagenesis process, the surrounding region was subcloned into a pLITMUST38 holding vector, by excising the fragment containing exon 14 (2762 bp) from the full length Flt1 pcDNA minigene (left) with *EcoRV* and *NheI*, and ligating into pL38 linearized with the same enzymes. The new exon 14 pL38 plasmid (right) was verified with restriction enzyme digests and sequencing of the ends, to confirm presence of entire fragment. Cloned portion is highlighted in both donor and acceptor vector.

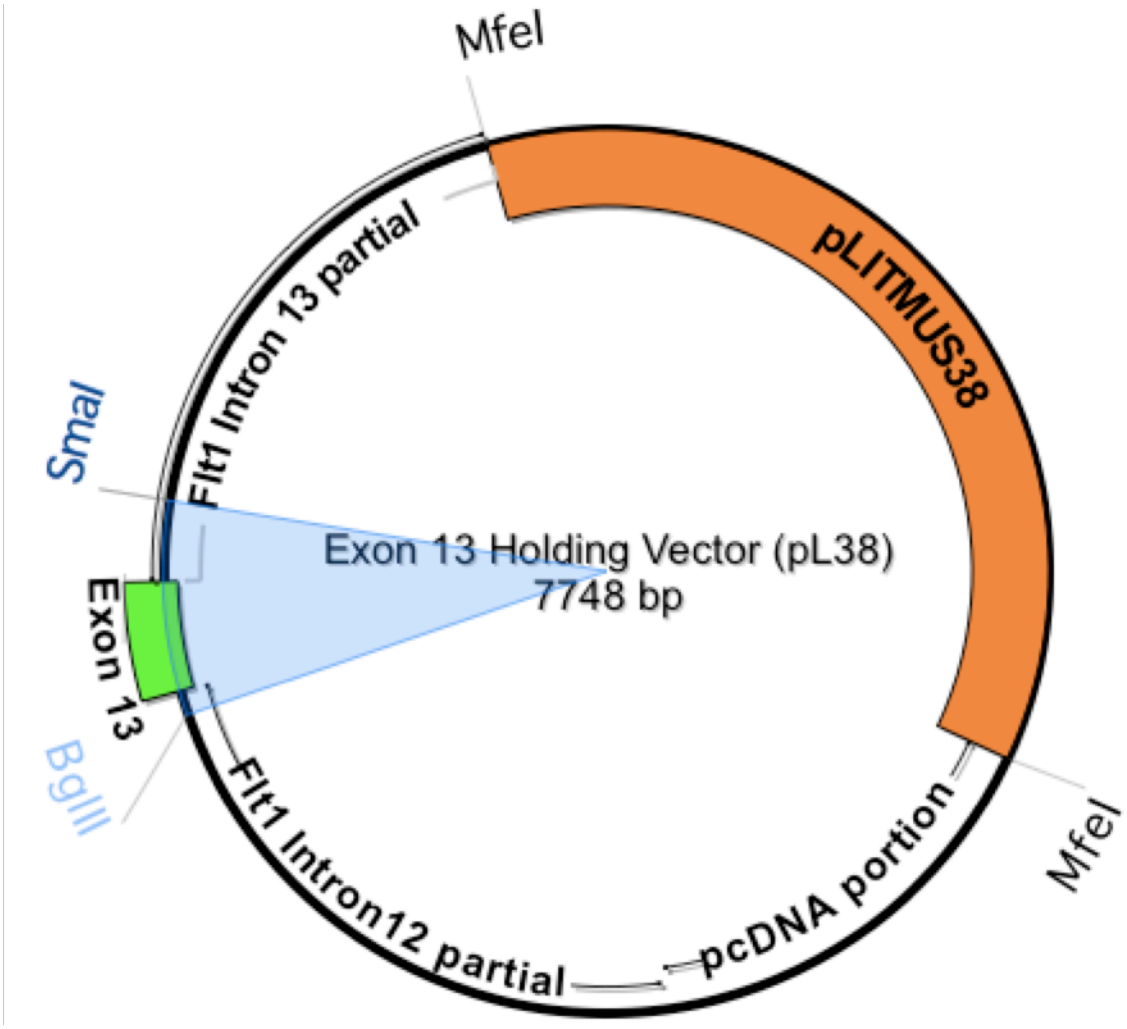


Figure 2.13: Creating Exon 13 Mutations. Using the Exon 13 pL38 holding vector, the exon 13 SNP and ESE mutations were created by amplifying plasmid with mutant primers (Table 2.4) in two successive PCR reactions. The primary PCR reaction produced two fragments that overlapped at the site of the mutation, while the secondary PCR reaction extended each end of these fragments to produce a long double stranded mutated DNA fragment. This mutated DNA fragment was then replaced in the holding vector by excising the exon 13 pL38 plasmid with *SmaI* and *BglII* (highlighted) to remove the WT sequence, and ligating in the amplified, mutated fragment (ends of fragment also cut with *SmaI* and *BglII*). This was done in two separate reactions, one for the SNP mutation and one for the ESE mutations.

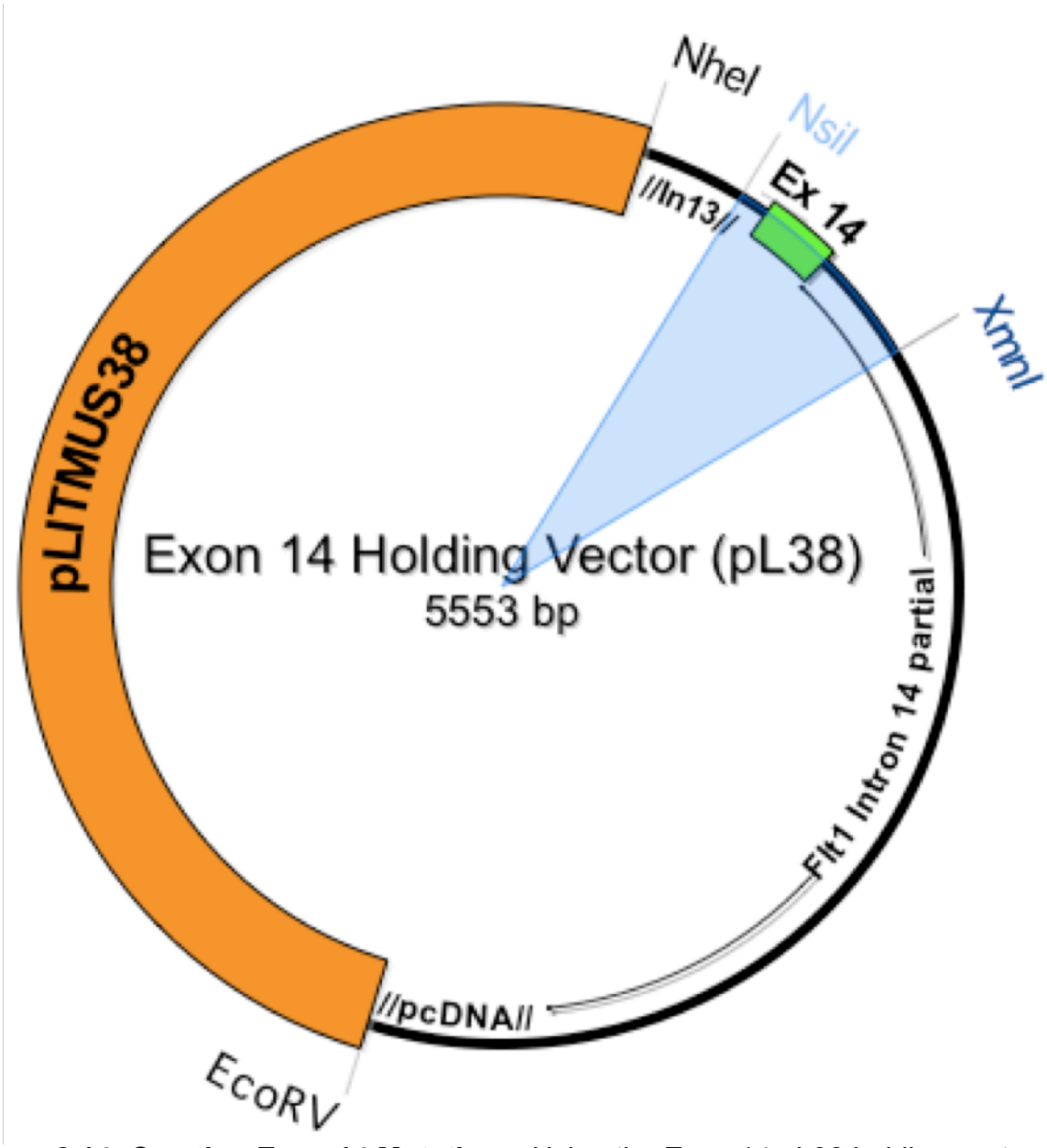


Figure 2.14: Creating Exon 14 Mutations. Using the Exon 14 pL38 holding vector, the exon 14 SNP and ESE mutations were created by amplifying plasmid with mutant primers (Table 2.3) in two successive PCR reactions. The primary PCR reaction produced two fragments that overlapped at the site of the mutation, while the secondary PCR reaction extended each end of these fragments to produce a long double stranded mutated DNA fragment. This mutated DNA fragment was then replaced in the holding vector by excising the exon 14 pL38 plasmid with *NsiI* and *XmnI* (highlighted) to remove the WT sequence, and ligating in the amplified, mutated fragment (ends of fragment also cut with *NsiI* and *XmnI*). This was done in two separate reactions, one for the SNP mutation and one for the ESE mutations.

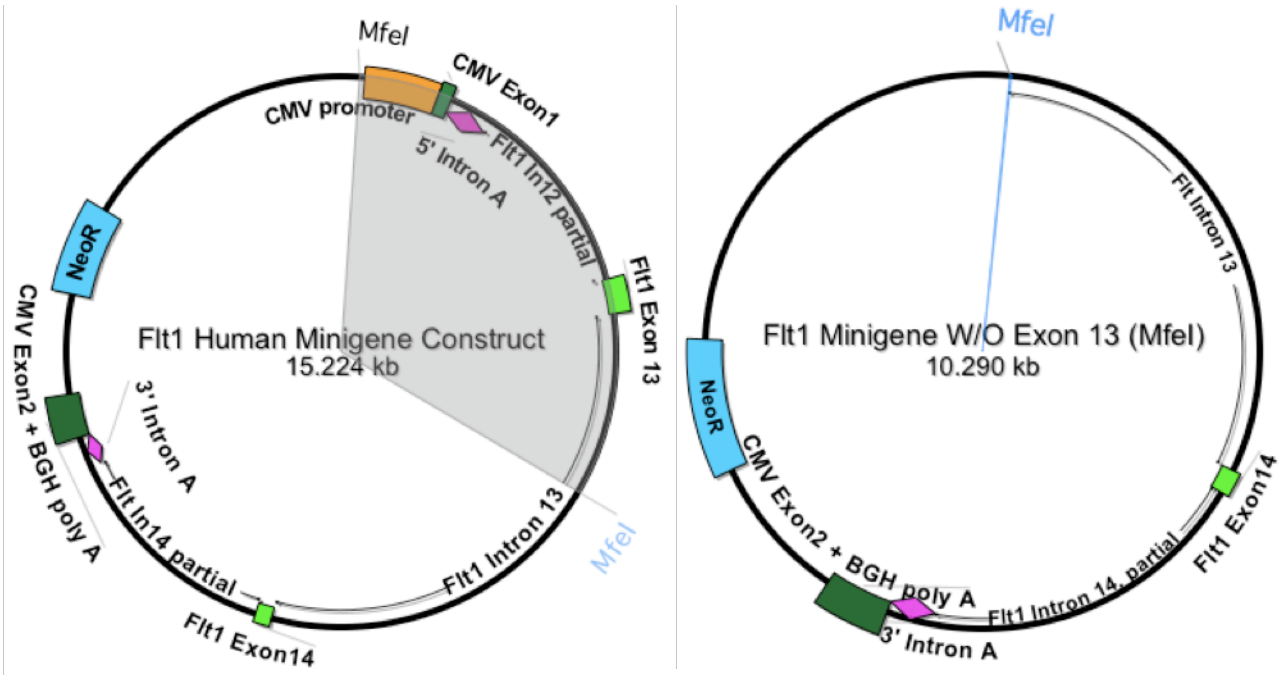


Figure 2.15: Extra step utilized to transfer mutated Exon 13 fragment from pL38 holding back into pcDNA vector. The *MfeI* fragment (4934 bases) containing the wild type exon 13 region was excised from the full length Flt1 minigene (left) to produce a plasmid lacking the wild type exon 13 and surrounding area. This 10.2Kb plasmid was then purified, linearized with *MfeI* and used as the acceptor plasmid for the mutated exon 13 SNP fragment and mutated exon 13 ESE fragment, to make full length pcDNA IntA mutated clones.

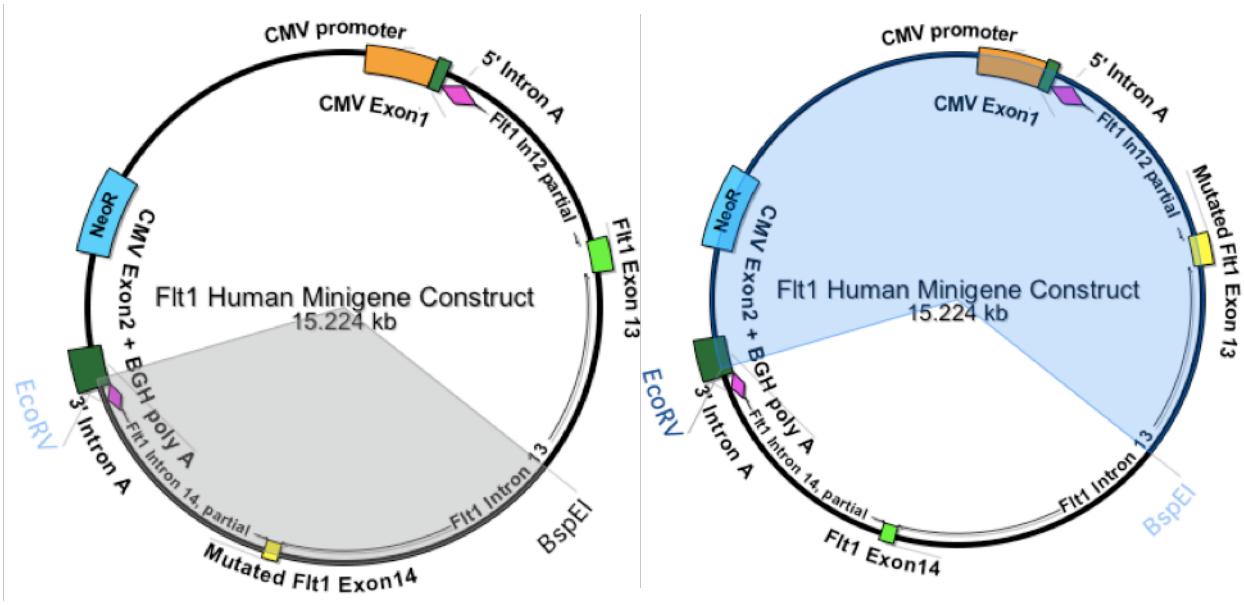
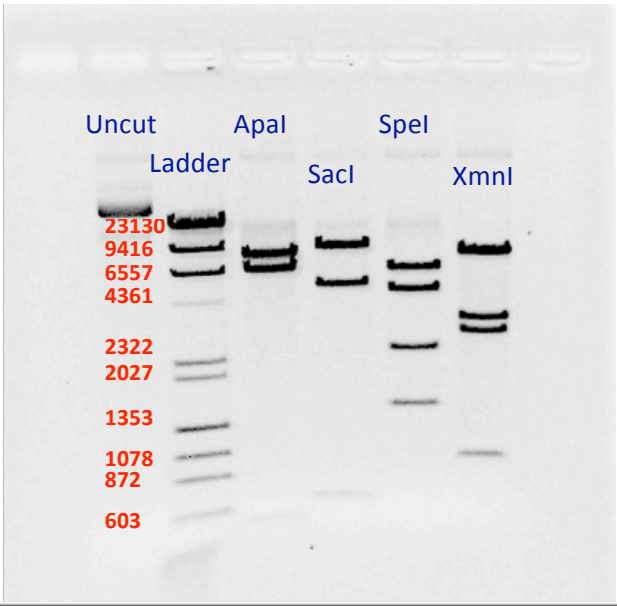


Figure 2.16: Making the double mutants. The exon 14 full length mutants (left) were used as "insert", excising the exon 14 region with using *BspEI* and *EcoRV* (highlighted portion is the excised fragment). The exon 13 full length mutants (right) were used as the "vector", linearizing with the same enzymes and discarding the exon 14 WT region (highlighted portion is the fragment used to receive the mutated exon 14 fragment). After ligating the exon 14 mutated region into the exon 13 mutated region, double clones were produced for SNP mutations and ESE mutations, respectively. (Mutated exon in each clone is shown in yellow, to distinguish from WT exon).



Name	Freq	Position
ApaI	4	3701 (13),3714 (449),4163 (6610),10773 (8152)
SacI	5	867 (642),1509 (8913),10422 (263),10685 (120), 10805 (5286)
SpeI	4	298 (6450),6748 (2447),9195 (1498),10693 (4829)
XmnI	4	7643 (961),8604 (2916),11520 (3382),14902 (7965)

Figure 2.17: Characterization of full length pcDNAIntA clones (Wild Type and all mutants) with restriction enzyme digests. A restriction enzyme digest was used to characterize each of the clones, using the 4 enzymes indicated above. The figure indicates expected fragment size for the full length human pcDNA IntA clones, and the gel depicts the fragment sizes, which appeared exactly as expected in all restriction digest for all mutants, as well as the WT. Fragment sizes of ladder are indicated in red on the gel picture.

Table 2.5: Nomenclature of clones in lab archives.

Clone Name	Description
BA 82	Wild Type
BA 366	Exon 13 SNP Mutant
BA 491	Exon 14 SNP Mutant
BA 555	Double SNP Mutant (both exons)
BA 500	Exon 13 ESE Mutant
BA 396	Exon 14 ESE Mutant
BA 548	Double ESE Mutant (both exons)

Figure 2.18: Sequencing of Exon 13 Mutant Clones to check for presence of mutations. Top Left: Exon 13 WT region (SNP highlighted), Top Right: Exon 13 non WT allele mutation in clone (highlighted). Bottom Left: Exon 13 WT (ESE region highlighted), Bottom Right: Exon 13 ESE mutations in clone (ESE region highlighted).

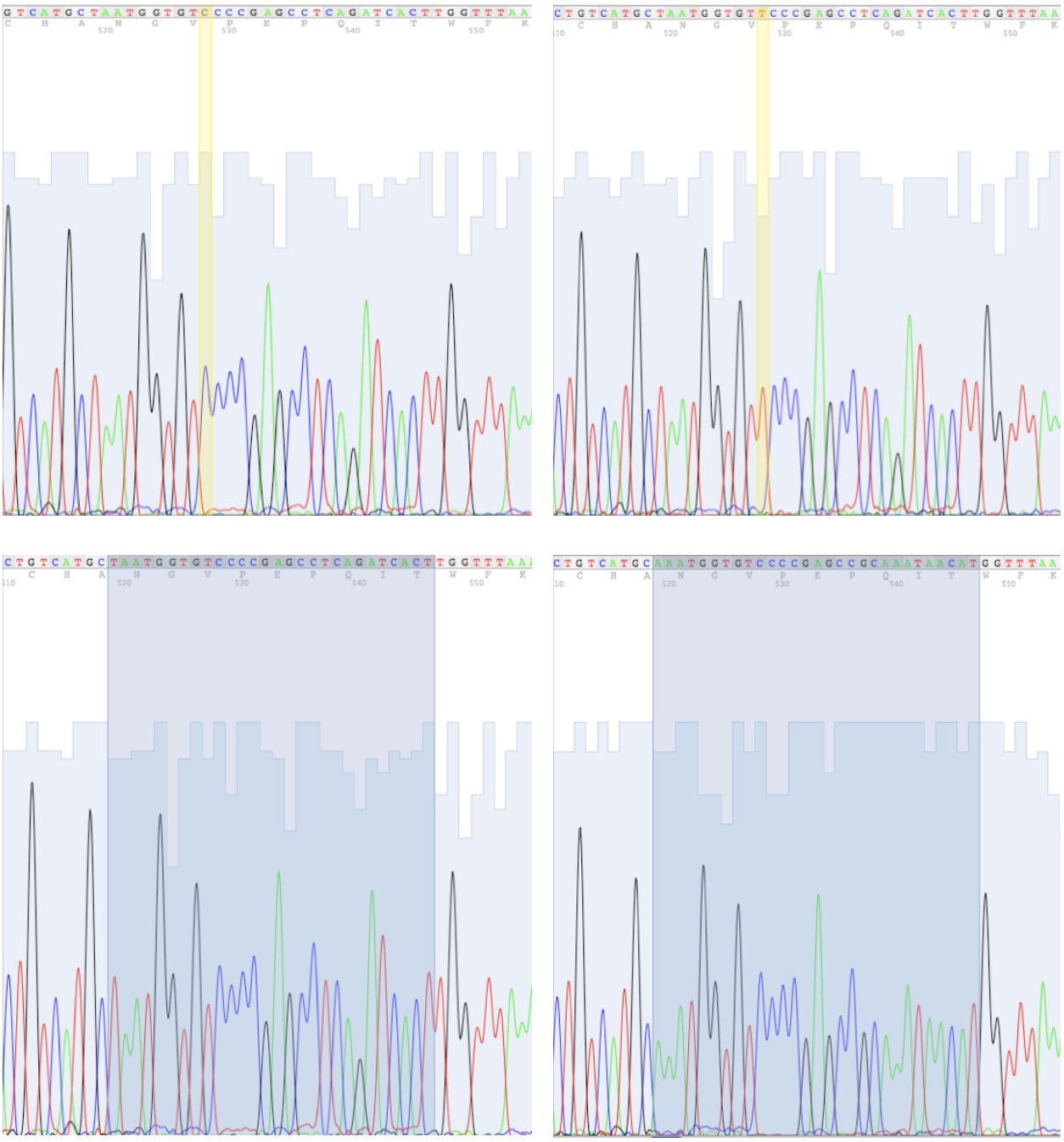
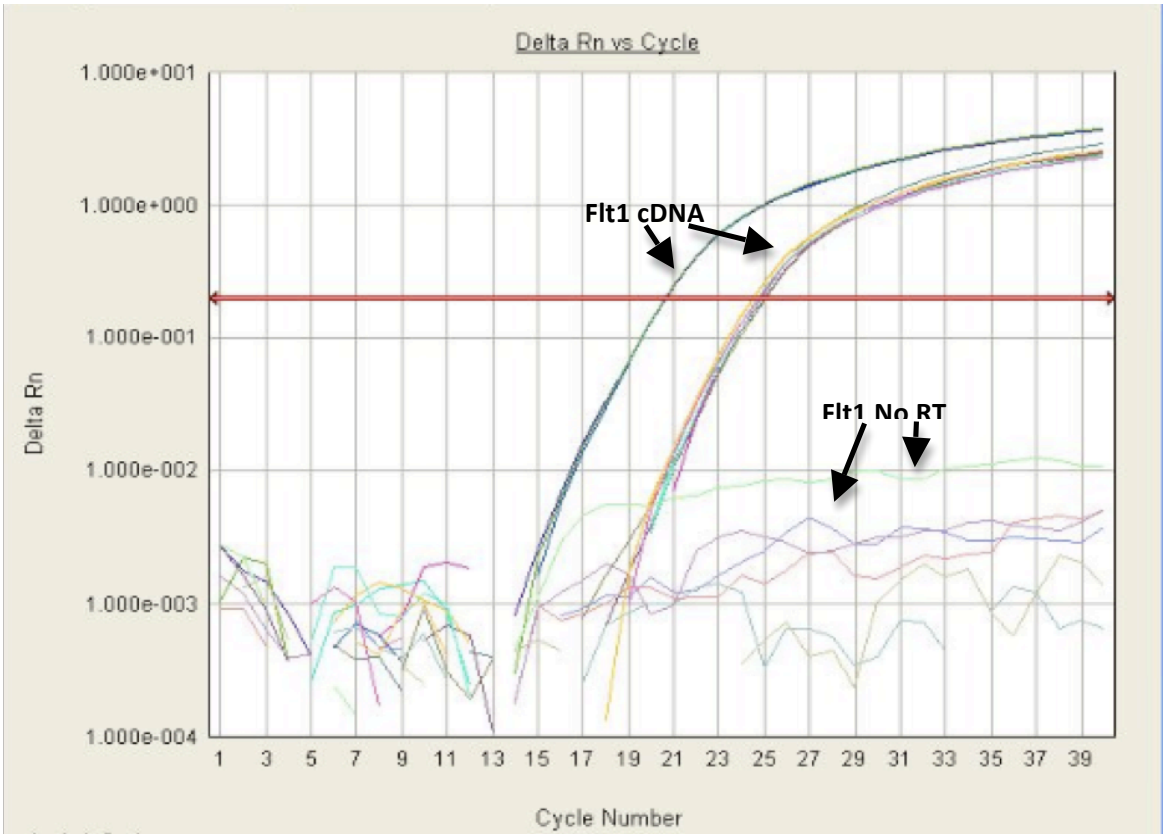


Figure 2.19: Sequencing of Exon 14 Mutant Clones to check for presence of mutations. Top Left: Exon 14 WT region (SNP highlighted), Top Right: Exon 14 non WT allele mutation in clone (highlighted). Bottom Left: Exon 14 WT (ESE region highlighted), Bottom Right: Exon 14 ESE mutations in clone (ESE region highlighted).



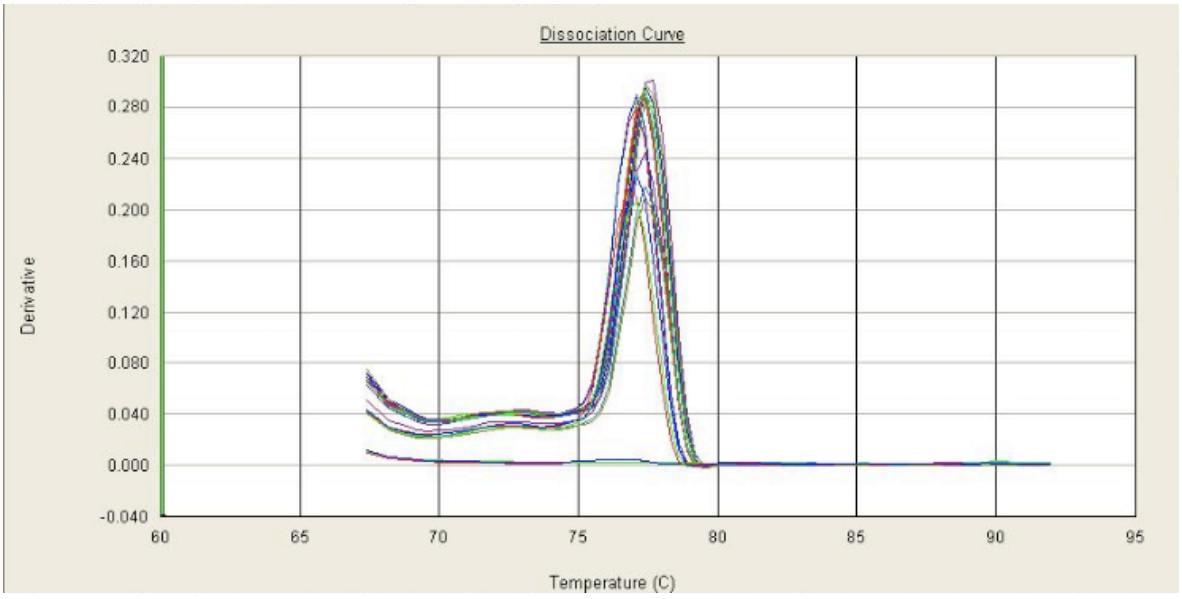
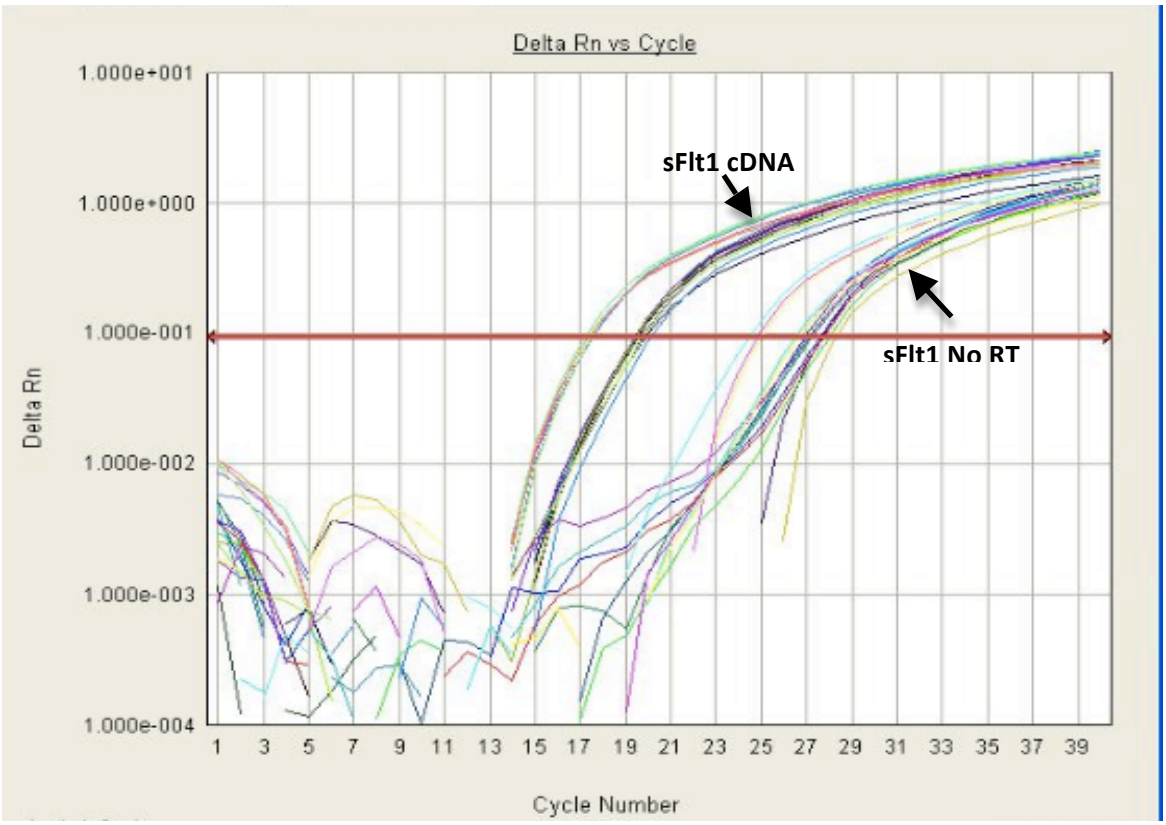


Figure 2.20. Flt1 qPCR examples. Top: Representative figure of Flt1 qPCR plots, with Flt1 cDNA curves and some Flt1 NoRT background annotated. Flt1 cDNA show clean amplification plots, while No RT samples are clearly in the background, as expected, since the Flt1 product does not exist in the plasmid product. Bottom: Representative figure of Flt1 dissociation curves, which depicts a very clean, single curve for all Flt1 cDNA, and no curve at all for No RT, as expected.



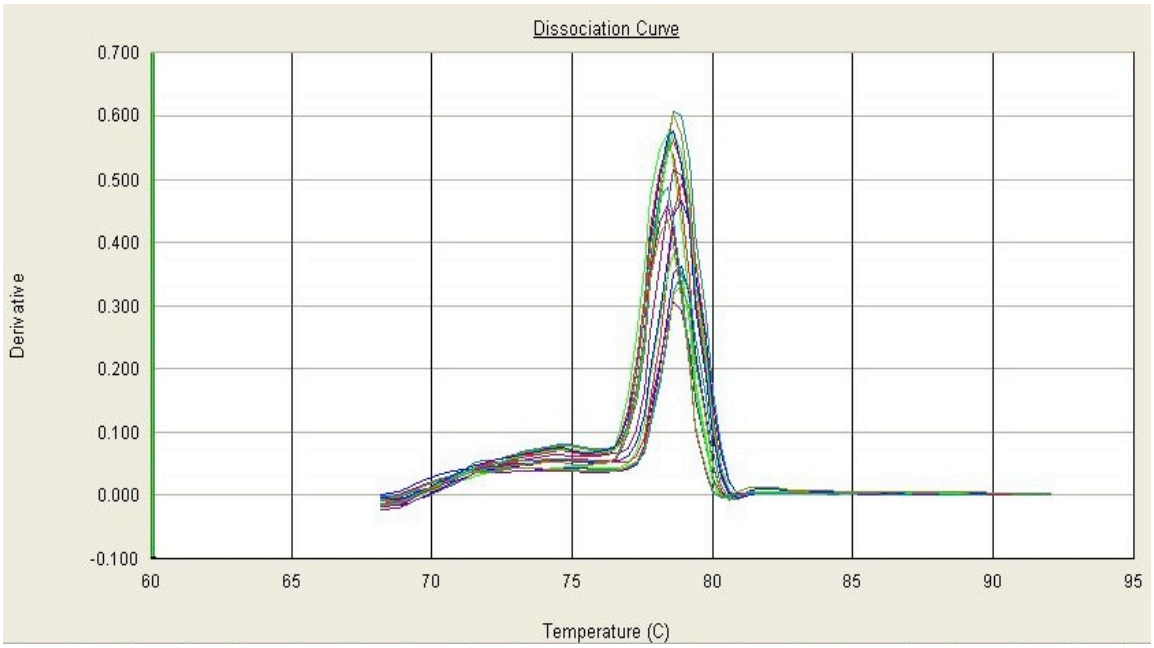


Figure 2.21. sFlt1 qPCR examples. Top: Representative figure of sFlt1 qPCR plots, with sFlt1 cDNA curves and sFlt1 NoRT background annotated. There is clear seperation between sFlt1 cDNA samples and No RT samples, as expected, since the sFlt1 RNA is much more abundant than the plasmid background. Bottom: Representative figure of sFlt1 dissociation curves, which depicts a single curve for all sFlt1 cDNAs and No RTs, thus implying one amplification product, as anticipated.

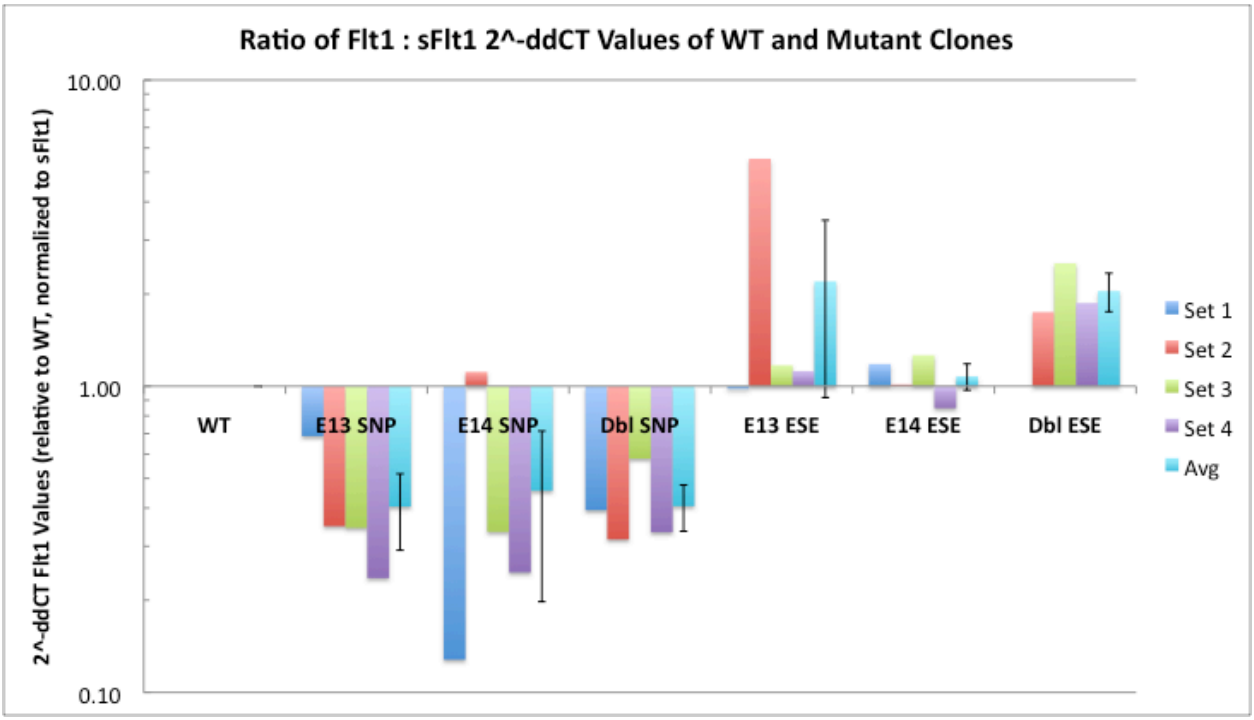


Figure 2.22: Change in the ratio of Flt1 to sFlt1 in mutant clones, as compared to Wild Type (four transfection sets and their average). As is evident in the figure, the ratio of Flt1:sFlt1 decreases with the presence of one or both SNP mutations. There is a decrease from WT expression of 40% and 46% of WT expression with a SNP in exon 13 or 14, respectively (standard error of mean (SEM) is 0.11 and 0.26, respectively), and a decrease of 41% from WT expression with a SNP in both exons (SEM = 0.07). The ESE mutations do not have a consistent effect on Flt1 and sFlt1 levels when present in either exon, as compared to WT expression levels. However, an ESE mutation in both exons increases the Flt1:sFlt1 ratios by slightly over 100% more than WT expression (SEM = 0.3). (2^-ddCT, plotted on a logarithmic scale; Average bars are teal; standard error bars are shown on “average” bars on top graph).

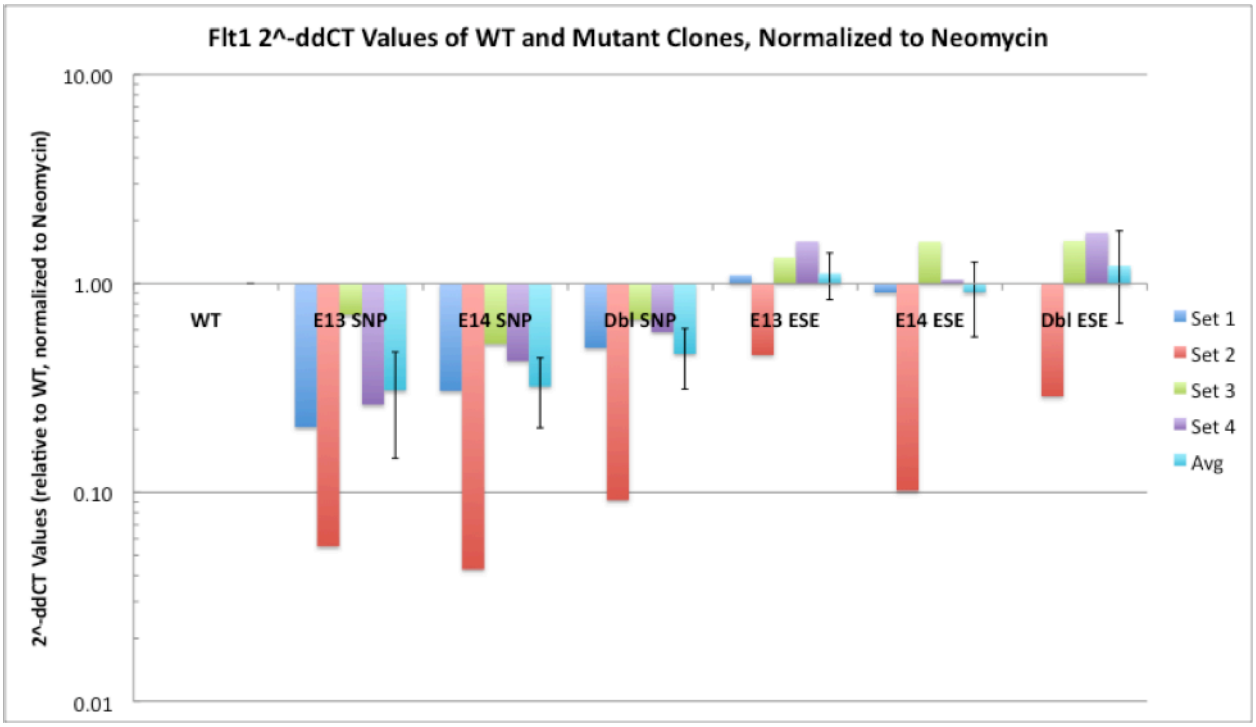


Figure 2.23: Change in the Flt1 levels, as compared to Wild Type; normalized to Neomycin (four transfection sets and their average). As is evident in the figure, the levels of Flt1 mRNA decrease to 30% and 32% of WT with the presence of the exon 13 and exon 14 SNP, respectively. When the SNP mutation occurs in both exons, there is a 46% decrease from WT expression in Flt1 (SEM = 0.07%). The ESE mutations do not have a consistent effect on Flt1 levels when present in either or both exons, as compared to WT expression levels. (2^-ddCT, plotted on a logarithmic scale; Average bars are teal; standard error bars are shown on “average” bars on top graph).

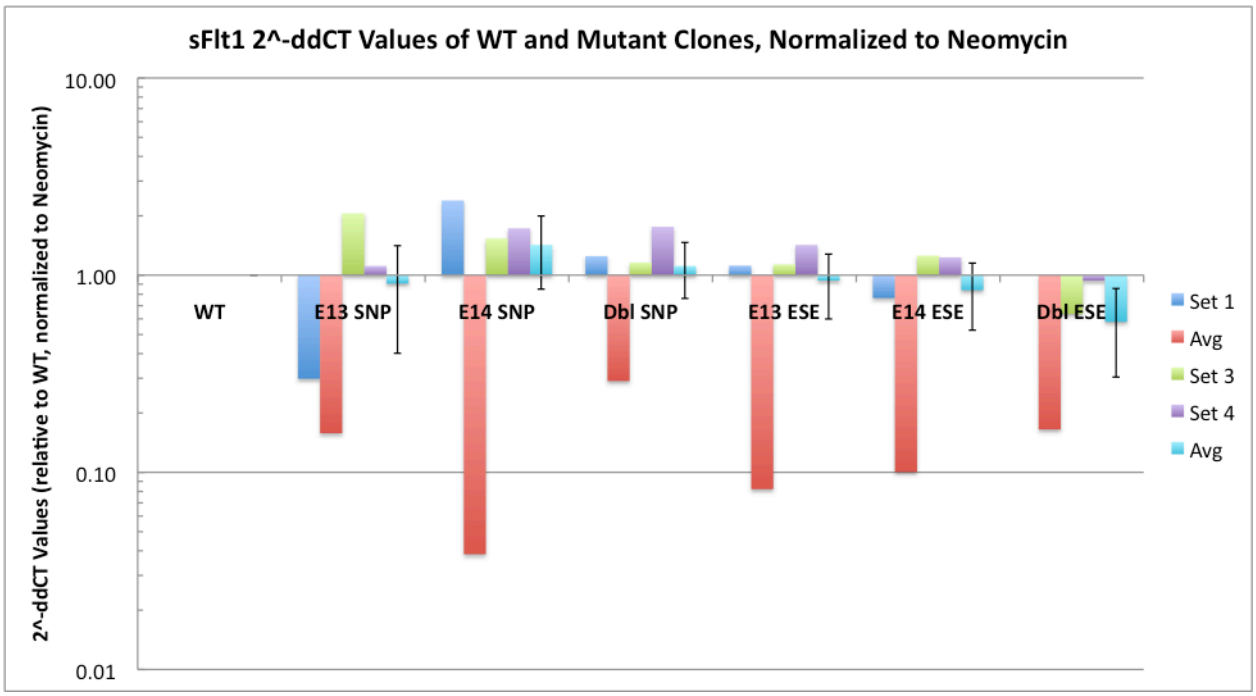


Figure 2.24: Change in the sFlt1 levels, as compared to Wild Type; normalized to Neomycin (four transfection sets and their average). As is evident in the figure, the levels of sFlt1 mRNA are not changed with the presence of the exon 13 SNP mutation or in the presence of both SNP mutations, but sFlt1 levels increase by about 42% in the presence of the exon 14 SNP mutation, when compared to wild type (SEM = 0.6). The ESE mutants ESEs in surrounding exons do not have a consistent effect on sFlt1 levels when present in either exon, as compared to WT expression levels. However, an ESE mutation in both exons decreases sFlt1 levels to 58% of WT expression (SEM=0.3). (2^-ddCT, plotted on a logarithmic scale; Average bars are teal; standard error bars are shown on “average” bars on top graph).

Chapter 3: Screen of Flt1 region for SNP Confirmation and Discovery

Abstract

Vascular endothelial growth factor (VEGF) is a potent angiogenic agent. VEGF activates its biologic responses through two cell-surface receptors, Flt1 and Flk1. In addition to the transmembrane form of Flt1, the Flt1 gene also encodes a secreted, truncated form of the receptor (sFlt1) translated from an mRNA in which a portion of intron 13 is preserved. sFlt1 retains high affinity for VEGF and thereby inhibits its angiogenic activity. Intron 13 contains important *cis* elements involved in sFlt1 mRNA formation. Here, we screened for polymorphisms in the human Flt1 gene, particularly SNPs at sites suspected to contain splicing or cleavage-polyadenylation signals that may influence Flt1 pre-mRNA processing and rates of Flt1 and sFlt1 expression. The NCBI SNP database contained 23 SNPs in the region of interest, one each in exons 13 and 14. Many of the SNPs have no population data, and the rest are incomplete. An independent human SNP screen (with a 115 sample size) of exons 13 and 14 with respective intron regions was conducted in order to confirm existing SNPs and search for new ones. Several of the reported SNPs were confirmed, but others had no incidence in this independent screen. Of those observed, incidence closely monitored levels reported by the databases, but a more thorough SNP screen of the entire Flt1 region is required. This is currently not a cost effective option, but may be so in the near future with the advent of more affordable high throughput next generation sequencing.

Introduction

The NCBI database (dbSNP) is the most comprehensive source for SNPs. This information is provided from a public effort of sequencing (HapMap) as well as private sequencing efforts (such as Celera) [83]. However, neither private nor public efforts have yet sequenced large numbers of individuals and as a result, lower frequency SNPs have likely gone undetected [83, 109]. More comprehensive efforts have been produced for 313 genes and only 2% of the SNPs identified were in dbSNP, which only reinforces the need for more thorough SNP investigation [83].

An independent SNP screen of the Flt1 gene region of interest as all currently identified SNPs have not been validated by "by multiple, independent submissions to the refSNP cluster". Some are double hits, which NCBI defines as alleles that have each been separately detected in at least two chromosomes. However, in the our region of interest in the Flt1 gene, most of the SNPs found have only been submitted by one source, at times two. Therefore, it is logical to assume that there have been overlooked SNPs. Thus, a screen of SNPs was envisioned to not only confirm the current SNP data, but also to identify any undiscovered SNPs in this region of the Flt1 gene.

Materials and Methods

Human subjects approval process

A screen of SNPs in the area of the Flt1 gene involved the collection of blood samples from volunteers. Consequently, the first step was ensuring compliance with all rules and regulations with the Institutional Review Board, which oversees any research involving human subjects to ensure that volunteers are treated ethically. The IRB proposal for this study (please see Appendix C for full proposal) was accepted, with approval code 08-636. This proposal included a consent form that volunteers were required to sign for entry into the study. As the study proceeded, several amendments were made to accommodate the changing needs of the study. These amendments involved the way volunteers were recruited as well as the sites of blood collection (Appendix C).

Volunteer recruitment, blood collection and DNA extraction

In order to ensure that the human genomic samples for this screen were distinct from those used in the HapMap project, the current study was performed using novel human genomic DNA collected from volunteers by way of blood samples. Volunteers were recruited via postings on email listservs across different departments, flyers hung around campus and word of mouth. Once a volunteer expressed interest, they either met with me or spoke on the phone and given a small background of the project and were walked through the contents of the consent form. They were asked to sign a consent form and provide ethnicity details for SNP classification purposes, and were then directed to donate blood.

Blood collection mainly occurred at the Virginia Tech Schiffert Health center, but other sites were utilized in which a licensed phlebotomist withdrew the blood, including Ohio State University-Zanesville. Samples were then collected and placed at 4°C for no more than 3 days, before proceeding with DNA extraction. In total, 4 mLs for 116 blood samples were collected, and 115 blood samples were used in the study. Table 3.1 depicts list of ethnicity distibution in samples.

DNA was extracted from 2 mLs of each blood sample using the QIAamp DNA Blood Midi Kit (Qiagen, Catalog #51183). Eluted DNA was checked with the NanoDrop to determine concentration. Total DNA extracted from samples ranged from a minimum of 96 µg to a maximum of 276 µg.

Choosing Regions to Screen and Primer Design

For the purpose of this study, we chose to study the two exons, 13 and 14, and surrounding parts of the introns to search for SNPs. Primers were designed around these areas and a few samples were chosen at random and amplification tested for PCR efficiency and sequencing efficiency. Once the quality of PCR amplification and sequencing was satisfactory, the amplification and sequencing steps were performed for all samples. Table 3.2 lists the primers used for each region, the sequence of each primer and the size of each amplified region.

Amplification and Sequencing

Through a sequence of trial and error, the following PCR reaction was developed: 0.2 uM of each primer, 100 -150 ng of DNA template and 25 µL of Taq PCR Master Mix, with nuclease-free water added to make up a 50uL reaction. Samples were run on the same standard PCR master mix as described in Appendix A. After PCR, 1 µL of each sample combined with 2 µL of Gel Loading Solution was run on a gel to visualize fragments. On every gel, a negative control (PCR reaction run alongside samples, but with no template) was also run to make sure there was no contamination of primers. Figure 3.1 shows an example of a gel from the exon 13 samples and a gel from the exon 14 samples.

Once it was determined that a PCR set had been successful, samples were purified from solution via desalting with the QIAEX II kit. Each DNA sample was eluted with 22 µL of nuclease-free water, 2 µL of which was used to measure the DNA concentration of amplified product. Once concentration was determined, 9 µL of each sample was mixed with 6 µL of sequencing primer (4.2 uM primer) in labeled tubes. Samples were placed on ice and shipped to Wake Forest Core Laboratory Facility (Appendix A) for sequencing. Once the sequences were received, they were aligned to the wild type sequence with the SeqMan Pro program of the Lasergene family of programs (DNASTAR).

Results

In the areas screened, there were seven reported SNPs, three of which were detected (Table 3.3). Total SNP data for detected SNPs in areas of exons 13 and 14 are summarized in Table 3.4.

Exon 13 SNPs: One out of two detected

There was no incidence of the exon 13 deletion SNP (rs72182275) but as is evident in Table 3.4, there were thirteen incidences of the exon 13 substitution SNP (rs17537350), with no data for seven samples. Twelve of these incidences were heterozygous, with only one homozygous for the non WT allele. Incidence was almost entirely Caucasian, with the exception of one Indian sample with the non WT allele (there was no data for seven samples). NCBI SNP database reports population data for Caucasian samples from two screens for this SNP: 6% non WT allele incidence, and

13% heterozygous in the first screen, 3% non WT allele incidence, and 4% heterozygous in the second screen, with 48 and 116 chromosomes screened, respectively. However, our current data indicates 11% total incidence, with 10% heterozygous. There is no information currently available for Indian populations, but the current data indicates a 1% incidence of heterozygous, with no homozygous incidence. NCBI database also reports incidence for African American populations (4% heterozygous, 2% incidence of non WT allele), whereas the current screen detected no non WT allele incidence; this is likely due to the low number of African American samples used in current screen (only two samples).

Exon 14 SNPs: None out of two detected

There were no incidences of the exon 14 substitution SNP (rs11843776), nor the exon 14 deletion SNP (rs66657484), although there was no data for 3 samples in the exon 14 region. There is no population data for the deletion exon 14 SNP, but reported population data for the substitution SNP is only in Sub-Sahran African populations (3% heterozygous, and 2% incidence of non WT allele in one screen, and 4% total incidence of non WT allele in a second screen, with no heterozygous data). The current lack of reported incidence may be explained by the fact that this screen only included 2 African American samples and no Sub-Saharan African samples.

Intronic SNPs: One detected in each of Introns 12 and 14, with high incidence

The two detected intronic SNPs had much higher incidence than the exonic SNPs. There were 62 incidences of the SNP in intron 12 (rs7983774), 57 in Caucasian samples, 3 in Indian samples, 1 in an African American and 1 in a sample with unknown ethnicity. There was no data for eight samples in this region. The NCBI SNP database reports three screens, each with differing percentages of incidence. For Caucasian samples, the first screen reports 47% non WT homozygous and 52% heterozygous incidence, 50% homozygous non WT and 33% heterozygous incidence and 67% total incidence of total non WT incidence (no homozygous/heterozygous data). The current data shows a much lower incidence of reported heterozygous incidence for Caucasian samples (only 7%), but agrees with the second screen for heterozygous incidence

(50%). For African American samples, the first screen and third screen in the NCBI SNP database list only Sub-Saharan African population (70% homozygous non WT, 28% heterozygous and 77% total non WT incidence with no homozygous/heterozygous, respectively) but the second screen reports 61% homozygous non WT incidence and 35% heterozygous incidence in the African American populations. The current screen indicates a much lower incidence (1% African American heterozygous incidence, with no incidence of homozygous WT allele), which is likely due to the small number of African American samples. There is no data reported for Indian samples but the current data indicates 3% heterozygous incidences, with no homozygous non WT incidences.

The substitution SNP detected intron 14 (rs3751395) showed the most common incidence of a non WT allele, found in 78 samples, with no data for three samples. Caucasian samples showed 72 incidences, Indian samples showed 5 incidences and African American showed 1 incidence. Three screens are reported on the NCBI SNP database for Caucasian samples: 17% homozygous non WT, 63% heterozygous, 32% homozygous and 58% heterozygous and 60% total non WT incidence (no homozygous/heterozygous data) for the three screens, respectively. The current screen indicates higher 30% homozygous and 35% heterozygous incidence for Caucasian samples. Only one of the three reported screens reports population data for African American samples (the first screen: 27% homozygous incidence and 40% heterozygous), and third reports incidence only for Sub-Saharan African populations (38% total incidence, with no homozygous/heterozygous data). The current data indicates 1% incidence of homozygous non WT allele and 1% heterozygous incidence for the African American population; lower numbers likely due to low number of African American samples in current screen. There is no data reported in the NCBI SNP database for Indian samples, but the current screen indicates a 3% homozygous non WT incidence and 2% heterozygous incidence.

Summarized Results

Percentage incidence of each detected SNP is summarized in Table 3.5, while incidence of each detected SNP, both homozygous alleles and heterozygous alleles are

shown in Figures 3.2 (Intron 12 SNP), 3.2 (Exon 13 SNP) and 3.4 (Intron 14 SNP). Figures 3.5-3.9 indicate the population distribution available for all reported SNPs.

Discussion

As interesting as these results, they are preliminary at best. For one, the samples are not ethnically diverse, so it is difficult to confirm the population distribution reported. In addition, ethnicity was self-reported by volunteers and may not accurately reflect ethnic background. As best as can be determined from the data, the SNPs that were confirmed appear to loosely agree in population distribution with NCBI SNP database reports (see Figures 3.5 – 3.9 for the population distribution for the intron 12 substitution SNP, the 2 substitution SNPs in exonic SNPs and the 2 intron 14 substitution SNPs – the remaining SNPs have no data reported).

In this study, I have attempted to screen SNPs in two exons in the Flt1 region, and was able to confirm the presence of several SNPs. However, this is only the beginning. A more thorough screen of SNPs in the Flt1 region is an extremely important step. Table 2.1 in Chapter 2 lists the SNPs that were available in the database, which were collected in 2005-2006. Three years later, the most updated list of SNPs in the Fl1 region has not changed by much: there are very few new SNPs, none with population data, and the ones previously reported do are still very lacking in population data (Table 3.6).

The Flt1 Intron 13 region is an extremely important regulatory region in Flt1/sFlt1 mRNA processing, as exemplified by previous work in the lab [2]. This region should be screened in order to confirm SNPs and discover undetected SNPs. However, the methods of SNP discovery utilized - DNA sequencing - is still very costly, and sequencing of a ~5000bp intron for over a hundred genomic samples would be an extremely costly undertaking. More cost-effective methods are available for SNP genotyping, ranging from custom targeted genotyping using TaqMan/qPCR probes to simple restriction enzyme methods. However, these methods are all limited in that they can be used to confirm existing SNPs, but cannot be used to search for new ones.

Thus, DNA sequencing for a full screen of the entire Flt1 region of Exon13-Intron13-Exon14 with conventional sequencing is not feasible at this time. However, with the advent of new high-throughput next generation sequencing technologies, such

as the ABI SoliD and Illumina, which rely on short-read technologies for the accurate discovery of both single-nucleotide and small insertion-deletion polymorphisms, this could be a very viable option in the near future. Sequencing accuracy is very high (less than 1% error rate), and cost per sample continues to drop, which may make this an affordable option very soon.

Table 3.1: Ethnicities of volunteers who donated blood for DNA extraction in SNP studies.

Ethnicity	Number of Samples	Percentage of Total
Caucasian	105 (3 Egyptian, 3 Hispanic, 3 Tunisian)	91%
Indian	6	5%
African American	2	2%
Vietnamese	1	1%
Unknown	1	1%

Table 3.2: Details of primers used to sequence Exon 13 and Exon 14, and respective surrounding areas.

Primer	Region	Orientation	Sequence	Fragment Size
BH 594	Exon 13	Sense (5'-3')	CCACCTGCTGTGTTTGCTCATT	676 bases
BH 595	Exon 13	Antisense (3'-5')	CTGCTATCATCTCCGAACTCA	
BH 597	Exon 14	Sense (5'-3')	GGAAGGAAACTAGCTAGCTAGG	548 bases
BH 598	Exon 14	Antisense (3'-5')	TGTTTGGGGCTCTATCAGCAAG	

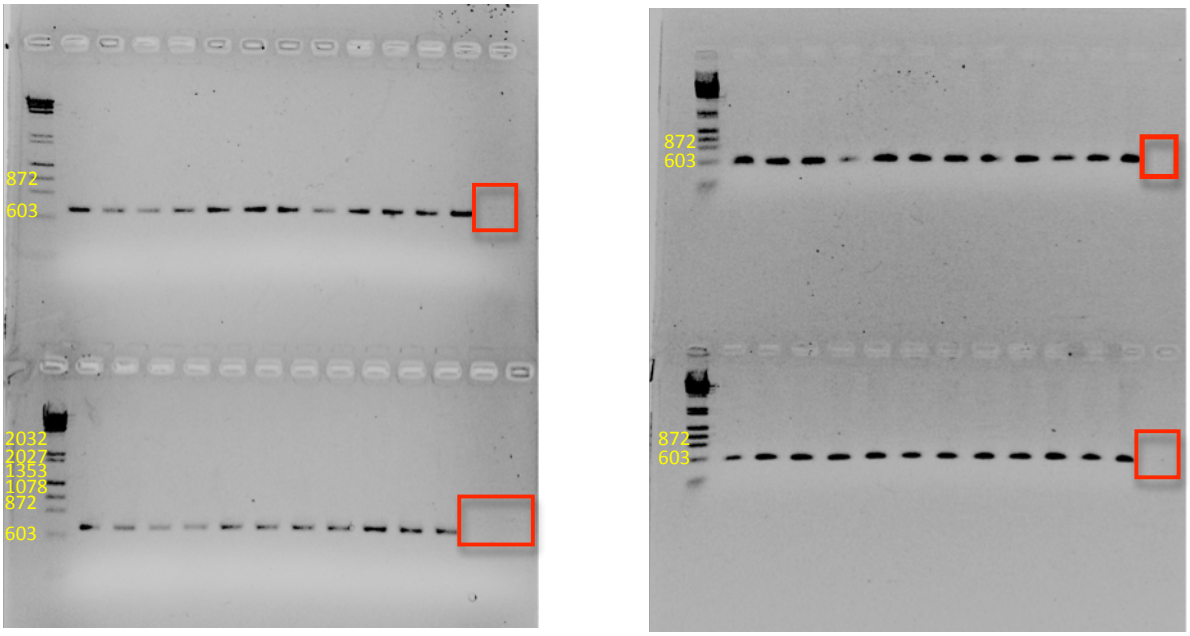


Figure 3.1: Examples of gels used to ensure correct amplification of human genomic DNA samples. After amplification of genomic samples, 1 µL of each sample combined with 2 µL of Gel Loading Solution was run on a gel to visualize fragments. Negative controls (PCR reaction run alongside samples, but with no template) were run alongside to make sure there was no contamination of primers (empty wells indicated by red boxes). Gel on left shows Exon 13 region, with a band size of 676 bases, while the gel of the left is the Exon 14 region, with a band size of 548 bases. Empty wells are the negative controls (red box). Size of marker (base length) indicated in yellow text beside marker band.

Table 3.3: Reported SNPs found in regions screened. SNPs for which a non WT allele was detected in this screen are indicated in red. Location of SNP on the gene indicated, as well the location in the two regions that were amplified and screened in the study. RS#s indicate NCBI identification numbers.

RS#	Location (Gene)	Location (Screen)	WT Allele	Non WT Allele
rs7983774	3' Intron 12	Exon 13 Region	C	T
rs17537350	Exon 13	Exon 13 Region	G	A
rs72182275	Exon 13	Exon 13 Region	A	Δ
rs66657484	5' Intron 13	Exon 13 Region	T	Δ
rs11843776	Exon 14	Exon 14 Region	C	T
rs66657484	Exon 14	Exon 14 Region	C	Δ
rs3751395	Intron 14 5'	Exon 14 Region	G	T
rs67203366	Intron 14 5'	Exon 14 Region	Δ	T
rs61763187	Intron 14 5'	Exon 14 Region	T	C

Table 3.4: List of SNPs detected in screen. WT: Wild Type; E: Egyptian, H: Hispanic; Amrc.: American; T: Tunisian; rs numbers: identification numbers for each SNP; - = No data; ? = undetermined base. Non WT alleles and heterozygous incidences are indicated in blue. The WT allele is only indicated if in disagreement with other primer; otherwise, an empty cell indicates wild type allele.

		Intron 12 SNP		**Exon 13 SNP**		**Intron 14 SNP**	
		rs7983774		rs17537350		rs3751395	
	PRIMER	5' to 3'	3' to 5'	5' to 3'	3' to 5'	5' to 3'	3' to 5'
WT		C		G		G	
NonWT		**T**		**A**		**T**	
Sample	**Ethnicity**						
1	Caucasian				-	**T**	**T**
2	Caucasian				-	**T**	**T/G**
3	Indian	**T/C**	**T/C**	**A/G**	**G/A**	**T**	**T/G**
4	Caucasian	**T/C**	**T/C**			**T/G**	**T/G**
6	Caucasian	-	C	-	-	**-**	**T/G**
7	Caucasian	**T/C**	-		-	**T**	
8	Caucasian	**T/C**	C				**-**
9	Caucasian	C	**T/C**			**T/G**	**-**
10	Caucasian-E	**T/C**	**T/C**				**-**
11	Caucasian-E	**T/C**	**T/C**			?	**T/G**
12	Caucasian-H	**T/C**	**T/C**			-	**-**
13	Caucasian	**T/C**	C				**-**
14	Caucasian		-		-	**T**	G
15	Vietnamese		-		-		**-**
16	Caucasian	**T/C**	**T/C**			**T/G**	**T/G**
17	Caucasian	**T/C**	**T/C**			**T/G**	**T/G**
18	Caucasian	**T/C**	**T/C**			**T**	G
19	Caucasian	**T/C**	C				**-**
20	Caucasian		-		-		**-**
21	Caucasian		**T/C**	**A/G**	G	-	G
22	Caucasian	**T/C**	**T/C**				**-**
23	Caucasian	**T/C**	C			**T/G**	G
24	Caucasian		-		-		**-**
25	Caucasian	**T/C**	**T**				**-**
26	Caucasian-T	**T**	**T/C**				**-**
27	Caucasian-T	**T/C**	C			**T/G**	**T**
28	Caucasian-T		-		-	**T**	**T/G**
29	Caucasian		-		-	**T/G**	**T/G**
30	Caucasian		-		-	**T/G**	**T/G**
31	Caucasian		-		-	**T/G**	**T/G**
32	Caucasian-E	**T/C**	**T**			?	G
33	Indian	**T/C**	**T/C**			**T**	**T/G**
34	Caucasian		C		**G/A**	-	-
35	Caucasian		-		-	-	**T/G**
36	Caucasian	**T/C**	**T/C**	**A/G**	**G/A**	**T**	-
37	Caucasian		C	**A/G**	**G/A**	**T**	**T**

		Intron 12 SNP		Exon 13 SNP		Intron 14 SNP	
		rs7983774		rs17537350		rs3751395	
	PRIMER	5' to 3'	3' to 5'	5' to 3'	3' to 5'	5' to 3'	3' to 5'
WT		C		G		G	
NonWT		**T**		**A**		**T**	
Sample	**Ethnicity**						
38	Caucasian	**T/C**	-		-		-
39	Caucasian			**A/G**	**G/A**	**T**	**T/G**
40	Caucasian	**T/C**	**T/C**	?		**T/G**	**T/G**
41	Caucasian	**T/C**	-		-	**T/G**	**T/G**
42	Caucasian	**T/C**	**T/C**				-
43	Caucasian	**T**	**T**				-
44	Caucasian		-		-	**T/G**	**T/G**
45	Caucasian		-		-	**T**	**T**
46	Caucasian	**T**	**T**				-
47	Caucasian	**T/C**	**T/C**				-
48	Caucasian		-		-	**T/G**	-
49	Caucasian		-		-		-
50	Caucasian	**T/C**	**T/C**				-
51	Caucasian	?	C			**T**	**T**
52	Caucasian		-		-	**T/G**	**T/G**
53	Caucasian	?	C			**T**	
54	Caucasian	**T/C**	-		-	**T**	**T**
55	Caucasian	**T/C**	-		-	-	**T/G**
56	Caucasian	**T/C**	-		-	-	
57	African Amr.		-		-	**T**	**T**
58	Caucasian	**T/C**	-	**A/G**	-		-
59	Caucasian		-	**A**	-	-	**T**
60	Caucasian		-		-		-
61	Caucasian	**T/C**	-	**A/G**	-	-	**T**
62	Caucasian		-		-	-	**T**
63	Unknown	**T/C**	-		-		-
64	Caucasian	-	-	-	-	**T**	**T**
65	Caucasian	?	-		-	**T/G**	**T**
66	Caucasian	**T/C**	-		-	**T/G**	**T**
67	Caucasian	**T?**	-		-	?	**T**
68	Caucasian	**T**	-		-	**T**	G
69	Caucasian	?	-	-	-	?	**T/G**
70	Caucasian	?	C			**T**	**T**
71	Caucasian	?	C			**T**	**T**
72	Caucasian		-		-	**T**	**T**
73	African Amr.	**T/C**	C			**T/G**	**T/G**
74	Caucasian	**T/C**	-	**A/G**	-	**T**	**T**
75	Caucasian	**T**	-		-	-	G
76	Caucasian	**T/C**	-	**A/G**	-	**T**	**T/G**
77	Caucasian		-		-	**T**	**T**
78	Caucasian					**T/G**	**T/G**
79	Caucasian		-		-	**T/G**	A

		Intron 12 SNP		Exon 13 SNP		Intron 14 SNP	
		rs7983774		rs17537350		rs3751395	
	PRIMER	5' to 3'	3' to 5'	5' to 3'	3' to 5'	5' to 3'	3' to 5'
WT		C		G		G	
NonWT		**T**		**A**		**T**	
Sample	**Ethnicity**						
80	Caucasian		-	**A/G**	-	**T**	**T**
81	Caucasian					**T**	**T**
82	Caucasian	**T/C**	**T**				-
83	Caucasian-H					**T**	**T**
84	Caucasian	**T/C**	-		-		
85	Indian		-		-	**T**	**T**
86	Caucasian		-		-	**T/G**	**T/G**
87	Caucasian	**T**	**T**			**T/G**	-
88	Caucasian		-		-		-
89	Caucasian		-		-	**T/G**	**T/G**
90	Caucasian	**T/C**	-		-		-
91	Caucasian		-		-	**T**	**T**
92	Caucasian	**T**	**T**				-
93	Caucasian		-		-	**T/G**	**T/G**
94	Caucasian		-	**A/G**	-	**T**	**T**
95	Caucasian	**T/C**	?			**T/G**	**T/G**
96	Caucasian		**T/C**			-	**T**
97	Caucasian	?	**T/C**			**T/G**	**T/G**
98	Caucasian	?	C			**T**	**T**
99	Indian	?	**T/C**			**T**	**T**
100	Caucasian-H	-	-	-	-	**T**	**T**
101	Caucasian	?	C			**T**	-
102	Caucasian	**T/C**	-		-	**T**	-
103	Caucasian	-	**T/C**	-		**T/G**	-
104	Caucasian	**T/C**	-	-	?		-
105	Caucasian		-		-	**T/G**	-
106	Caucasian	**T/C**	**T/C**			**T/G**	-
107	Caucasian	**T/C**	**T/C**			**T/G**	-
108	Caucasian	**T/C**	C				-
109	Caucasian		-		-	**T/G**	-
110	Caucasian	**T/C**	-	-	-	**T**	-
111	Caucasian	**T/C**	**T/C**	-		**T/G**	-
112	Caucasian	?	**T/C**			-	-
113	Indian	?	-		-		-
114	Indian	?	-		-	**T**	-
115	Caucasian	-	-	-	-		-
116	Caucasian	-	-	-		**T**	-

Table 3.5: Percentage Incidence of each confirmed SNP. Non WT incidence is indicated both as total number of samples with non WT allele/heterozygous sample, as well as a percentage of each in total sample set. Ethnicity of SNP incidence in each ethnic group is indicated, as a percentage of sample set.

SNP	Non WT: % (#)	Homozygous Non WT: %(#)	Heterozygous: %(#)
Intron 12 (rs7983774)	58% of total (62)	7 % Caucasian (7) 0 % Indian (0) 0 % African American (0) 0 % Unknown (0)	46 % Caucasian (50) 3 % Indian (3) 1 % African American (1) 1 % Unknown (1)
Exon 13 (rs17537350)	13% of total (13)	1 % Caucasian (1) 0 % Indian (0)	10% Caucasian (11) 1% Indian (1)
Intron 14 (rs3751395)	71% of total (78)	30 % Caucasian (33) 3 % Indian (3) 1 % African American (1)	35 % Caucasian (39) 2 % Indian (2) 1 % African American (1)

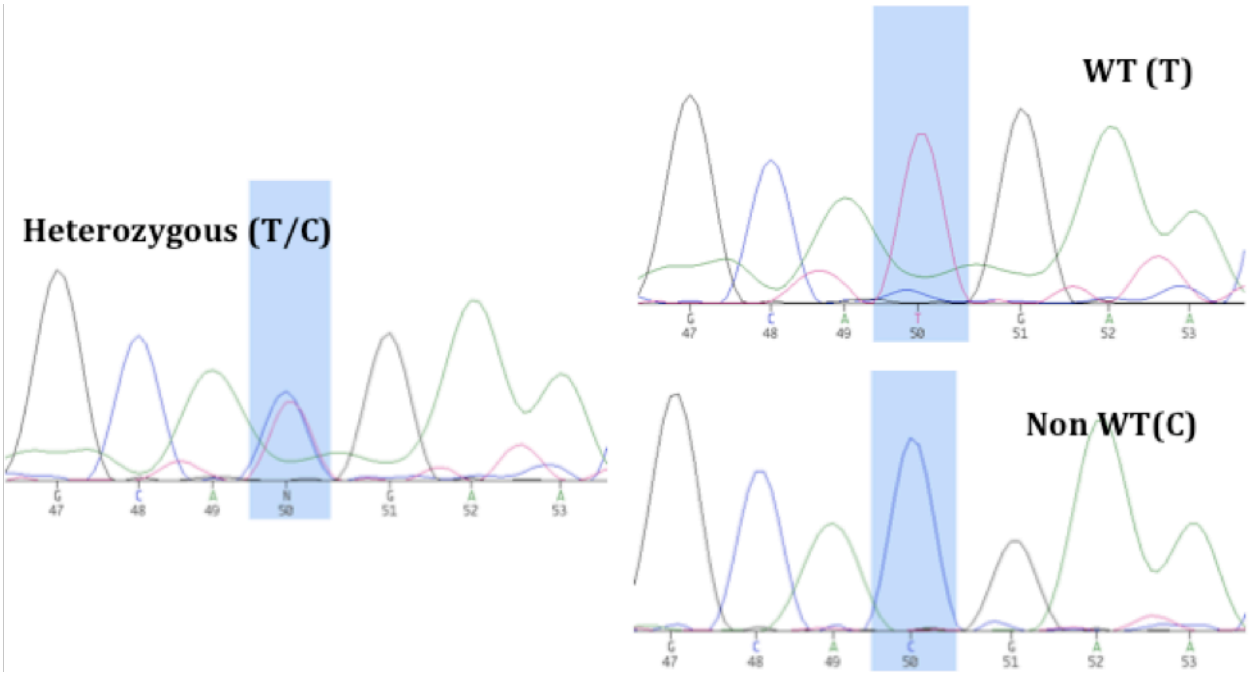


Figure 3.2: Illustration of SNP found in Intron 12 (rs7983774). Picture on left illustrates heterozygous incidence, while right illustrates individual incidences of alleles, both WT (top right) and non WT (bottom right). Location of SNP is highlighted.

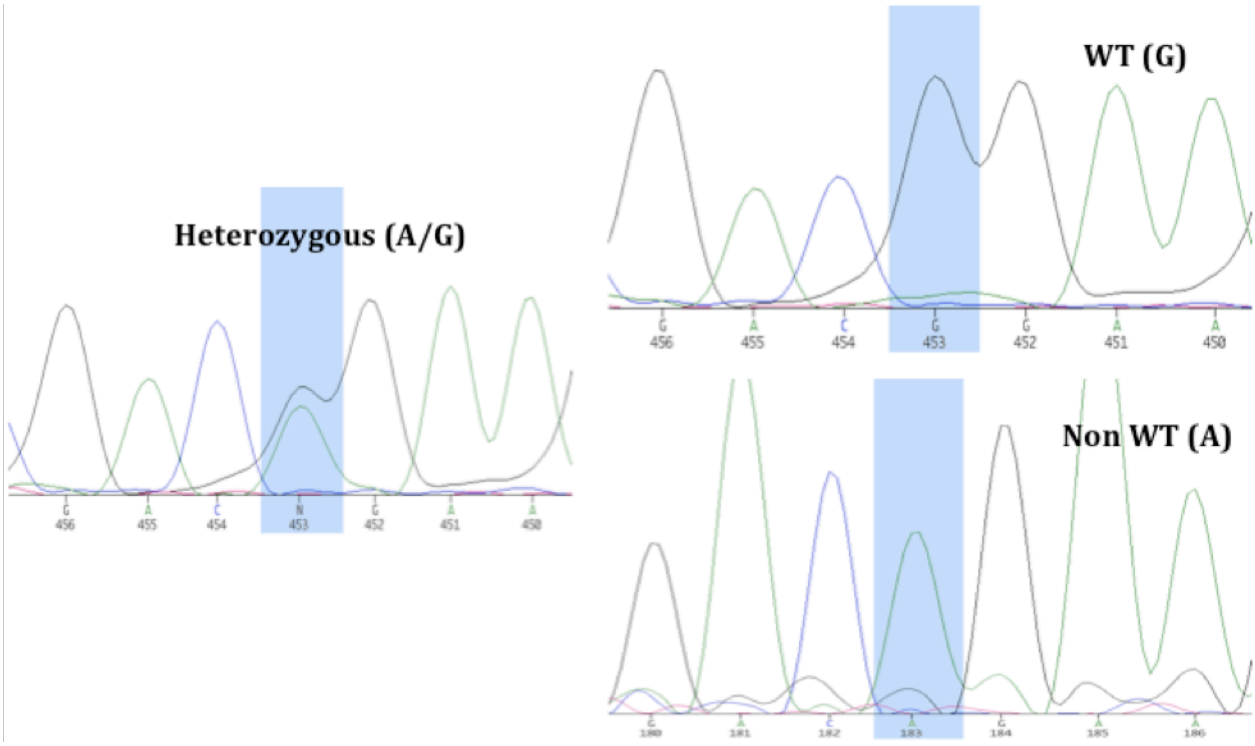


Figure 3.3: Illustration of SNP found in Exon 13 (rs17537350). Picture on left illustrates heterozygous incidence, while right illustrates individual incidences of alleles, both WT (top right) and non WT (bottom right). Location of SNP is highlighted.

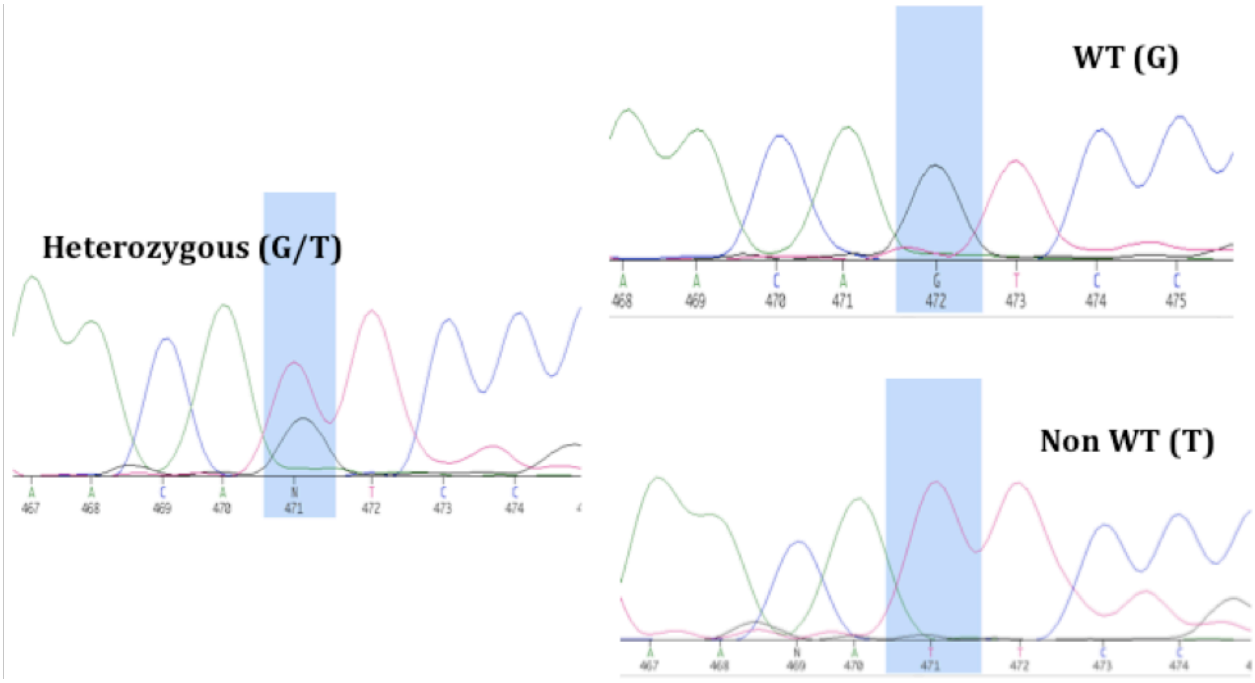
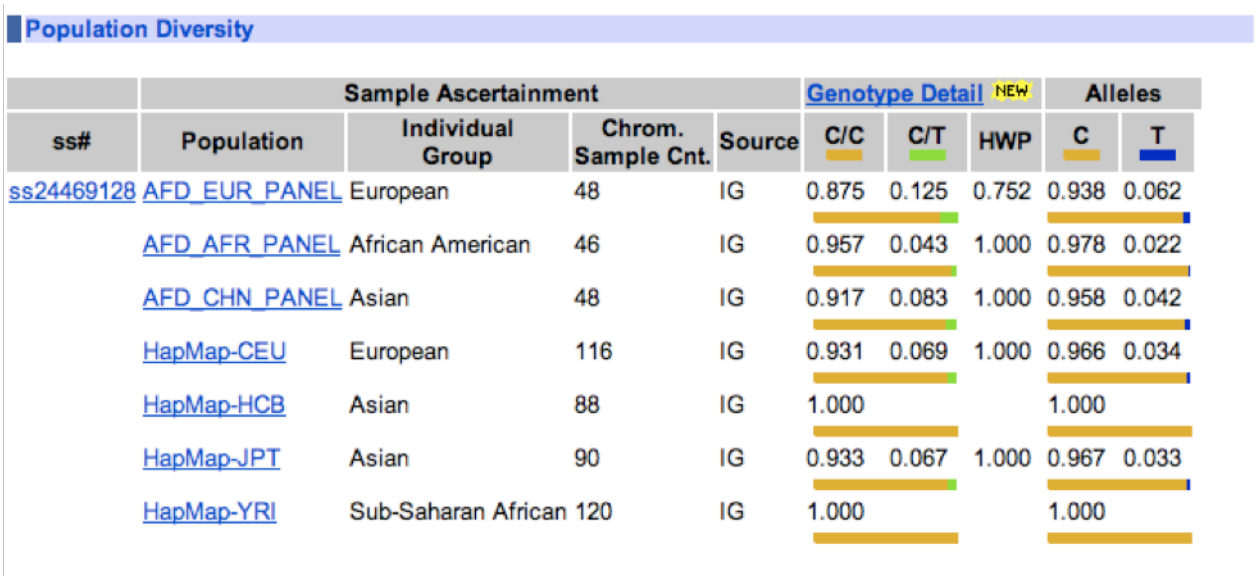


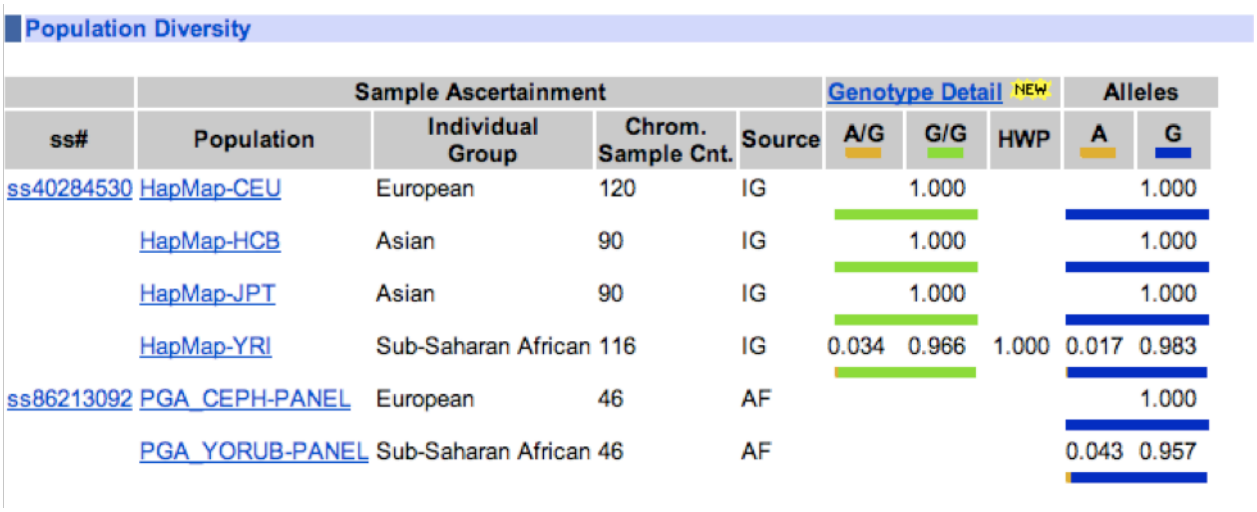
Figure 3.4: Illustration of SNP found in Intron 14 (rs3751395). Picture on left illustrates heterozygous incidence, while right illustrates individual incidences of alleles, both WT (top right) and non WT (bottom right). Location of SNP is highlighted.



Population Diversity

	Sample Ascertainment				Genotype Detail NEW			Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	C/C	C/T	HWP	C	T
ss24469128	AFD_EUR_PANEL	European	48	IG	0.875	0.125	0.752	0.938	0.062
	AFD_AFR_PANEL	African American	46	IG	0.957	0.043	1.000	0.978	0.022
	AFD_CHN_PANEL	Asian	48	IG	0.917	0.083	1.000	0.958	0.042
	HapMap-CEU	European	116	IG	0.931	0.069	1.000	0.966	0.034
	HapMap-HCB	Asian	88	IG	1.000			1.000	
	HapMap-JPT	Asian	90	IG	0.933	0.067	1.000	0.967	0.033
	HapMap-YRI	Sub-Saharan African	120	IG	1.000			1.000	

Figure 3.5 [3]: Population Diversity for substitution SNP reported in exon 13 (rs17537350). Abbreviations: HWP = Hardy Weinberg Equilibrium Data; ss= Submitted SNP reference number; Chrom Sample Count= the number of chromosomes that the submitter used as the denominator in computing estimates of allele frequencies. IG = Individual Genotype.



Population Diversity

	Sample Ascertainment				Genotype Detail NEW			Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	A/G	G/G	HWP	A	G
ss40284530	HapMap-CEU	European	120	IG		1.000			1.000
	HapMap-HCB	Asian	90	IG		1.000			1.000
	HapMap-JPT	Asian	90	IG		1.000			1.000
	HapMap-YRI	Sub-Saharan African	116	IG	0.034	0.966	1.000	0.017	0.983
ss86213092	PGA_CEPH-PANEL	European	46	AF					1.000
	PGA_YORUB-PANEL	Sub-Saharan African	46	AF				0.043	0.957

Figure 3.6 [3]: Population Diversity for substitution SNP reported in exon 14 (rs11843776). Abbreviations: HWP = Hardy Weinberg Equilibrium Data; ss= Submitted SNP reference number; Chrom Sample Count= the number of chromosomes that the submitter used as the denominator in computing estimates of allele frequencies. IG = Individual Genotype, AF = Allele Frequency

	Sample Ascertainment				Genotype Detail NEW				Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	A/A	A/G	G/G	HWP	A	G
ss12233220	HapMap-CEU	European	120	IG	0.017	0.517	0.467	0.025	0.275	0.725
	HapMap-HCB	Asian	88	IG	0.068	0.568	0.364	0.150	0.352	0.648
	HapMap-JPT	Asian	86	IG	0.163	0.512	0.326	0.752	0.419	0.581
	HapMap-YRI	Sub-Saharan African	120	IG	0.017	0.283	0.700	0.655	0.158	0.842
ss24021493	AFD_EUR_PANEL	European	48	IG	0.167	0.333	0.500	0.251	0.333	0.667
	AFD_AFR_PANEL	African American	46	IG	0.043	0.348	0.609	1.000	0.217	0.783
	AFD_CHN_PANEL	Asian	48	IG	0.083	0.750	0.167	0.020	0.458	0.542
ss86213091	PGA_CEPH-PANEL	European	46	AF					0.326	0.674
	PGA_YORUB-PANEL	Sub-Saharan African	48	AF					0.229	0.771

Figure 3.7 [3]: Population Diversity for SNP reported in Intron 12 (rs7983774). Abbreviations: HWP = Hardy Weinberg Equilibrium Data; ss= Submitted SNP reference number; Chrom Sample Count = the number of chromosomes that the submitter used as the denominator in computing estimates of allele frequencies. IG = Individual Genotype, AF = Allele Frequency

	Sample Ascertainment				Genotypes	Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	HWP	C	T
ss86213093	PGA_CEPH-PANEL	European	46	AF			1.000
	PGA_YORUB-PANEL	Sub-Saharan African	46	AF		0.022	0.978

Figure 3.8 [3]: Population Diversity for first substitution SNP reported in Intron 14 (rs61763187). HWP = Hardy Weinberg Equilibrium Data. Abbreviations: HWP = Hardy Weinberg Equilibrium Data; ss= Submitted SNP reference number; Chrom Sample Count= the number of chromosomes that the submitter used as the denominator in computing estimates of allele frequencies. AF = Allele Frequency

	Sample Ascertainment				Genotype Detail NEW				Alleles	
ss#	Population	Individual Group	Chrom. Sample Cnt.	Source	A/A	A/C	C/C	HWP	A	C
ss24469076	AFD_EUR_PANEL	European	48	IG	0.208	0.625	0.167	0.251	0.521	0.479
	AFD_AFR_PANEL	African American	46	IG	0.348	0.391	0.261	0.317	0.543	0.457
	AFD_CHN_PANEL	Asian	48	IG	0.250	0.625	0.125	0.200	0.562	0.438
ss43491806	HapMap-CEU	European	118	IG	0.102	0.576	0.322	0.150	0.390	0.610
	HapMap-HCB	Asian	88	IG	0.364	0.500	0.136	0.752	0.614	0.386
	HapMap-JPT	Asian	90	IG	0.422	0.444	0.133	1.000	0.644	0.356
ss4937422	JBIC-allele		1492	AF					0.686	0.314
ss71643270	HapMap-CEU	European	118	IG	0.102	0.576	0.322		0.390	0.610
	HapMap-HCB	Asian	88	IG	0.364	0.500	0.136		0.614	0.386
	HapMap-JPT	Asian	90	IG	0.422	0.444	0.133		0.644	0.356
ss86213094	PGA_CEPH-PANEL	European	42	AF					0.405	0.595
	PGA_YORUB-PANEL	Sub-Saharan African	40	AF					0.625	0.375

Figure 3.9 [3]: Population Diversity for second substitution SNP reported in Intron 14 (rs3751395). Abbreviations: HWP = Hardy Weinberg Equilibrium Data; ss= Submitted SNP reference number; Chrom Sample Count= the number of chromosomes that the submitter used as the denominator in computing estimates of allele frequencies. IG = Individual Genotype, AF = Allele Frequency

Table 3.6: Screen of SNPs in Flt1 region in 2009, updated from 2006. Three years later, SNP data in Flt1 region is still lacking population data. Exonic SNPs highlighted in yellow. †Listed as Double Hit in SNP Browser but not NCBI, *Also listed as part of a 2bp insertion (GA/TT, rs35409953 with no frequency data), Val = Validation status (Nt= Not Validated, D = Double hit, V = Validated), Source: N = NCBI, S = SNP Browser, B = Both

SNP ID	SNP	WT	Val	Source	African		Caucasian		Chinese		Japanese	
rs17086623	C/T	C	D	N	1.00	0.00	1.00	0.00	0.92	0.08	N/A	
rs34483960	-/A	-	Nt	N	No Data							
rs17086620	A/G	A	D	B	N/A		0.93	0.07	0.99	0.01	1.00	0.00
rs7983774	A/G	G	D	N	0.16	0.84	0.28	0.72	0.35	0.65	0.42	0.58
rs17537350	C/T	C	D	B	N/A		0.97	0.03	1.00	0.00	0.97	0.03
rs72182275	-/A	A	Nt	N	No Data							
rs2296283	A/G	G	D	B	0.07	0.93	0.58	0.42	0.63	0.37	0.52	0.48
rs2296284	A/G	G	D	B	0.14	0.86	0.27	0.73	N/A		N/A	
rs58053169	-/A	A	Nt	N	No Data							
rs60862114	-/AA	AA	Nt	N	No Data							
rs67400923	-/A	A	Nt	N	No Data							
rs5802479	-/A	A	Nt	N	No Data							
rs17086618	A/G	A	Nt	B	0.96	0.04	1.00	0.00	1.00	0.00	1.00	0.00
rs4771246	C/T	C	D	B	1.00	0.00	0.95	0.05	1.00	0.00	0.97	0.03
rs9551465	A/T	A	D	B	0.93	0.07	0.41	0.59	0.37	0.63	0.48	0.52
rs17086617	C/T	T	D	B	0.15	0.85	0.28	0.72	0.35	0.65	0.44	0.56
rs7337610	C/T	T	D	B	0.18	0.82	0.35	0.65	0.36	0.64	0.48	0.52
rs61707719	-/T	T	Nt	N	No Data							
rs66532169	-/A	A	Nt	N	No Data							
rs58793540	-/A	A	Nt	N	No Data							
rs11400597	-/A	-	Nt	N	No Data							
rs56065910	T/A	A	Nt	N	No Data							
rs7325639	C/T	C	D	B	No Data							
rs60370048	T/A	A	Nt	N	No Data							
rs17626077	G/T	G	D	B	0.92	0.08	0.94	0.06	1.00	0.00	1.00	0.00
rs35630136	-/G	-	Nt	N	No Data							
rs3088111	C/T	T	D	B	0.08	0.92	0.00	1.00	0.00	1.00	0.00	1.00
rs3751397	A/T	A	D	B	0.12	0.88	0.58	0.42	0.58	0.42	0.53	0.47
rs3751396	A/G	A	D	B	1.00	0.00	1.00	0.00	0.99	0.01	0.97	0.03
rs11618105	A/C	A	D	B	No Data							
rs7319123	A/C	A	D	B	0.18	0.82	0.00	1.00	0.00	1.00	0.00	1.00
rs71748205	-/CA	CA	Nt	N	No Data							
rs73453252	C/T	T	D	B	No Data							
rs3794404	A/C	A	D	B	1.00	0.00	1.00	0.00	0.99	0.01	0.98	0.02
rs11843776	A/G	G	D	B	0.02	0.98	0.00	1.00	0.00	1.00	0.00	1.00
rs66657484	-/G	G	Nt	B	No Data							
rs67203366	-/A	-	Nt	B	No Data							
rs61763187	A/G	A	Nt	B	0.98	0.22	1.00	0.00	N/A		N/A	

rs3751395	**A/C**	**C**	D	B	**0.63**						
rs3751395	**A/C**	**C**	D	B	N/A	**0.39**	**0.61**	**0.61**	**0.39**	**0.64**	**0.36**
rs34113415	**-/T**	-	Nt	N	No Data						
rs11347871	**-/A**	**A**	Nt	N	No Data						
rs11294603	**-/A**	**A**	Nt	N	No Data						
rs12871995	**A/C**	**C**	Nt	N	No Data						
rs12872498	**A/C**	**C**	Nt	N	No Data						
rs7140067	**C/T**	**C**	D	B	**0.91**	**0.09**	**0.97**	**0.03**	**1.00**	**0.00**	**1.00**	**0.00**
rs13378470	**C/T**	**C**	Nt	N	No Data						
rs6491281	**C/T**	**T**	Nt	N	No Data						
rs13378467	**C/T**	**C**	Nt	N	No Data						
rs6491280	**A/G**	**G**	D†	B	**0.99**	**0.01**	**1.00**	**0.00**	**1.00**	**0.00**	**1.00**	**0.00**
rs9551464*	**A/T**	**A**	Nt	N	No Data						
rs9551463*	**G/T**	**G**	Nt	N	No Data						

Chapter 4: ESE Mutator program: "The EXONerator"

Abstract

Splicing of introns is an essential step in the pre-mRNA processing events that must occur in genes to form a functional messenger RNA (mRNA) competent for nuclear export and translation. The reactions involved in intron removal are catalyzed by the "spliceosome," which interacts with consensus *cis*-acting splice signals in the pre-mRNA as well as other supplementary elements such as exonic splice enhancers (ESEs) that allow the spliceosome to accurately differentiate between pseudosites and legitimate splice sites. ESEs act as binding sites for SR proteins, which facilitate exon inclusion. Understanding how ESEs function and the factors involved in exon definition and inclusion will not only lead to a better understanding of alternative splicing, but also to a deeper understanding of aberrant splicing in diseases. There are many computational tools available that can help predict a potential ESE site within an exon, and its expected strength of binding to SR proteins, such as the ESEfinder. In this work, we have created a PERL program (The EXONerator), which automates the process of mutating predicted ESE sequences to silence their effect. This greatly simplifies the generation of coding sequence with silenced ESEs, allowing for further study of the effects of ESEs in a coding region of interest. As an example, the Flt1 coding region of exon 14 was run through The EXONerator, and 24 predicted ESEs in this region were silenced. In conclusion, The EXONerator is an extremely valuable tool that can be tailored for use as required for experimental design and greatly simplifies the process of silencing ESEs for testing their functional significance. As more information about SR proteins surfaces, the PERL coding can be easily modified to include information about other RNA binding proteins as they become available.

Introduction

In higher eukaryotes, most genes are made up of DNA sequence that is divided into exons and introns[110]. Splicing of introns is an important step in the pre-mRNA processing events that must occur in such genes, to form a functional messenger RNA (mRNA) competent for nuclear export and translation [110, 111]. The reactions involved in intron removal are catalyzed by the "spliceosome," a large ribonucloprotein complex formed by a stepwise assembly of the snRNPs U1, U2, U4/U6 and U5, as well as non-snRNP proteins on the pre-mRNA [111]. This construction of the spliceosome call for interactions with *cis*-acting splice signals in the pre-mRNA, such as the 5' splice site, the 3' splice site, the polypyrimidine tract and the branchpoint sequence [111]. While these consensus signals are necessary in eukaryotic pre-mRNA processing, they must often supplemented by other *cis* elements within the exon or intron that can act to enhance or repress the splice signals[111]. These supplementary elements are necessary due to the introns containing sequences with high conservation to splice-sites (also known as "pseudosites") but are not utilized in the splicing reaction [110]. Collectively, these elements function in allowing the spliceosome to accurately differentiate between pseudosites and legitimate splice sites [111]. Identifying these supplementary *cis*-elements and their method of function is an important step in understanding not only pre-mRNA processing, but alternative splicing as well.

One such supplementary *cis*-element is sequences known as exonic splice enhancers (ESEs), which help the spliceosome recognize the few valid splice sites found among the plethora of pseudosites [111, 112]. In mammalian genes, ESEs were initially identified as purine-rich sequences that allowed utilization of flanking splice sited via an association with proteins from the SR family [4]. The non-snRNP splicing factors, which include proteins of the SR family and SR-related proteins, contain one or more RS domains (domains rich in alternating arginine and serine residues) [4]. The SR proteins and SR-related proteins differ in both structure and function, but their shared feature of RS domains makes them essential in supporting the formation of the spliceosome by assembling on the ESE sequences and recruiting the spliceosome

components [4]. SR proteins also appear to concurrently interfere with the function of adjacent splicing silencers, further enhancing splice enhancer signal at ESE sites [113].

ESEs have an effect on disease, whereby point mutations in an exon lead to splicing defects but these point mutations are not present in a splice signal nor do they create a cryptic splice site [4]. Some examples of genetic diseases that are a direct result of mutations in ESE sequences include spinal muscular atrophy, Becker muscular dystrophy and frontotemporal dementia linked to chromosome 17 [114-116].

While such diseases are a result of a mutation forcing a change in splice site choice, we still have a limited understanding of the factors that are involved in alternative splice site selection [117]. Therefore, an understanding of how ESEs function and the factors involved in exon definition and inclusion will not only lead to a better understanding of alternative splicing, but also to a deeper understanding of aberrant splicing in diseases. To this end, there are many computational tools available that can help predict a potential ESE site within an exon, and its expected strength of binding to SR proteins. One such program is the freely-available ESEfinder, a valuable web based tool used to predict the location and strength of ESE motifs for four SR proteins [1]. The ESEfinder sets a “threshold” value for each protein, and any value above the threshold is considered a high scoring ESE, with scores varying based on the sequence in question [1]. ESEfinder has been used to predict the presence of ESEs in a number of genes involved in disease [1]

The ESEfinder program has a variety of limitations [1]. The first is that the threshold values and scores for ESE motifs and are somewhat arbitrary: while the actual motifs were determined via functional SELEX (‘systematic evolution of ligands by exponential enrichment), the threshold values are somewhat arbitrary as they are based on statistical analysis and empirical data. In addition, the program only identifies ESE motifs for four human SR proteins (SF2/ASF, SC35, SRp40 and SRp55), as ESE motifs have not been identified for other SR proteins and could not be incorporated [1]. Perhaps the most important limitation is that the strength of an ESE motif is computationally predicted, so if the ESEfinder predicts ESEs in certain sequences, these sequences do not necessarily constitute an actual ESE in the context of the gene. This is partly because there may be adjacent silencer elements that can counteract the

effects of ESEs, which the ESEfinder does not take into account. Therefore, there needs to be a way to test these predicted ESE sequences in the context of the gene. An attempt to mutate ESEs in a very small portion of the Flt1 exons 13 and 14 was a laborious task (when taking into account the need to preserve amino acid sequence), and it was not completely effective (not all ESEs were silenced in the regions of exons 13 and 14). Thus, the rationale for developing The EXONerator was to simplify the task of mutating predicted ESE sequences, while preserving the amino acid sequence, which can then be further tested in the lab for functional significance.

The EXONerator allows an investigator to enter a coding sequence of interest, and generate a coding sequence with all ESEs with scores below threshold, predicting their silencing. These resulting sequences with silenced ESEs could then be tested in the context of the gene, wherein their effect can be tested.

Materials and Methods

Program Design

The PERL programming language was chosen (suggested by Dr. Brett Tyler) to write The EXONerator for a variety of reasons. PERL is easy to learn, can be written in a simple text editor and is easy to run on the three major operating systems (Windows, Macintosh and Unix) currently in use on desktop computers [118]. In addition, PERL was well suited to our goals, as it is a has very high text and pattern matching capabilities (matching the entered DNA sequence to the recognized ESE motifs) and manipulating strings of text (mutating the DNA sequences to silence the ESEs) [118]. These features made PERL the most logical choice of programming language for creating The EXONerator.

Envisioning how the program would work was straightforward: the program needed to accept DNA sequences in a format that would suit the investigator, via simple copy-and-paste of a plain text sequence or by uploading a plain text document with the sequence. Once the DNA had been entered, the program would then identify ESEs and score them, compare the score against the threshold and then proceed to mutate the DNA in a way that would not only bring the score of an ESE below threshold, but also preserve the amino acid sequence so as not to introduce any change in the DNA except

for silenced ESEs. Lee Falin assisted in envisioning the elements needed for the program to function correctly.

Coding

As mentioned, the code for The EXONerator was written in the PERL programming language, with help from Jihane Najdi in panning out the more complicated subroutines required to achieve the goals intended. The first thing the program does is check to make sure it is a coding (open reading frame) sequence. If the entered sequence contains nucleotides not found in DNA (C,T,A or G) or its length is not evenly divisible by three, an error message will be generated.

Once it has been established that the entered sequence is in the form of a DNA sequence, this sequence is divided up into overlapping blocks of 18 bases. The length of the block was chosen because of the motif lengths: a single base change in a motif of 8 bases long (the longest motif for the SR proteins considered) can affect 7 bases upstream and 7 bases downstream, an entire block of bases must be taken into account when mutating DNA. The most efficient way to do so was found by way of trial and error: the DNA string is divided up in blocks of 18 bases, with a shift of 9 bases downstream each time so that each preceding block overlapping with the next block with the last 12 bases (Figure 4.1). Once the first block is changed, the modified block replaces the original, so that the second block includes any modifications made, and so on.

Each block of 18 bases is then passed through several subroutines, where the score is calculated for each protein motif sequentially. Each motif is checked at each position for protein motifs in the 18-base block, shifting one position each cycle (Figure 4.2 illustrates the logic). Once the program gets to a position where there are too little bases to make up a motif, it ceases checking that particular protein (for example, SF2 has a 7 base motif, so when there are six bases left in a block, the program ceases to check SF2). These unchecked bases are re-checked within the next block, since they overlap, so it does not present a problem of overlooked ESE sequences.

The program starts at position 1, computes the scores for all 4 SR proteins and checks these scores against each threshold. If a score is found that is above threshold,

the motif is then changed according to the rules for mutating while preserving the amino acid sequence. The first rule is that only a codon coding for the same amino acid can be substituted in place of the codon in question. The second rule is based on the position for which the program is computing scores. While scores are computed based on the motif length (6, 7 or 8 bases long, depending on the SR protein), mutations are made only if the position it is looking at is the beginning of a codon triplet: if it is in the second or third base of a codon, it will ignore these 1-2 bases when mutating, and move on until it finds the first base that constitutes a codon. Any change in the DNA results in a cycle so that all proteins are re-checked for motif scores, to make sure that silencing one ESE will not introduce an ESE for another protein at the same position.

Please refer to Appendix D for the complete PERL code of The EXONerator.

Results: Validation and Testing

The EXONerator was tested extensively with several coding sequences, and it is an extremely valuable tool. In all cases, comparing the generated sequence with silenced ESEs with the original entered sequence shows no change in the coding sequence.

For example, the Flt1 exon 14 coding sequence (total length of 144 bases) was run through the EXONerator. This sequence included a total of 24 ESE motifs with varying strengths: 4 for SF2, 9 for SC35, 9 for SRp40 and 2 for SRp55 (Figure 4.3).

The first generated sequence resulted in silencing of 19 of these ESEs, with only 5 ESEs remaining in the sequence (Figure 4.4). This generated sequence, with only 5 ESEs was run through The EXONerator again, which resulted in the silencing of all 5 ESEs, but introducing 3 new ones that were not previously present (Figure 4.5). This second generated sequence, with only 3 ESEs was run through for a third and final time, with the generation of a sequence with no ESEs (Figure 4.6).

In total, 28 bases (out of 144) were changed, but there was no difference in amino acid composition of entered sequence and generated sequence (Figure 4.7). Therefore, The EXONerator has achieved the goals intended: effectively silence ESEs in a given coding sequence, while preserving the amino acid sequence.

Comparing the manually mutated ESE sequences surrounding exon 13 and exon 14 SNPs (Chapter 3) with that mutated via The EXONerator showed a much higher efficiency. In the region surrounding the exon 13 SNP, manual mutation of ESE sequences silenced only three out of five predicted ESEs, with five base changes being made, but utilization of The EXONerator for this region silenced all five predicted ESEs, with a total of eight bases changed (Figure 4.8 and 4.9). In the region surrounding the exon 14 SNP, manual mutation of ESE sequences silenced only three out of five predicted ESEs, with five base changes being made, but utilization of The EXONerator for this region silenced all five predicted ESEs, with a total of six bases changed (Figure 4.10 and 4.11).

Discussion

The EXONerator is an extremely valuable tool that can be used to silence predicted ESEs in a small fragment of a coding sequence, or all predicted ESEs in a coding sequence of an exon. In essence, it can be tailored for use as required by the experimental design. One can enter any coding sequence of interest, no matter how short or long, and receive silenced ESEs in that particular region.

There are several caveats to keep in mind while using The EXONerator. The first concerns the output of the sequence, which may differ from run to run, even with the same input. This is due to the program mutating the DNA sequence to lower the score below threshold in a random manner: it will identify all nucleotides that contribute to a score below threshold which can substitute for the amino acid in a particular position, and then choose one of the amino acids at random. This was necessary in the code, as a change in the motif of one protein may bring that particular ESE below threshold, but can introduce an ESE for a different protein. The program will then cycle infinitely between these two proteins, mutating one to silence one ESE while introducing another, only to mutate again to silence the newly introduced ESE and reinstate the original ESE. Randomizing the choice of nucleotides allows the program to avoid this infinite loop. While the program is highly efficient at mutating the DNA to silence all ESEs, sometimes the first output will not silence all ESEs due to the randomization in the

choice of nucleotides. The investigator has two choices: to either run the program again with the sequence of interest, or to enter the generated sequence of the program (i.e. the sequence of interest with most ESEs silenced), and run again. In most cases, the remaining ESEs will then be silenced.

Another thing to keep in mind (especially in the case of larger exons) is that all ESEs in the sequence may not be silenced, no matter how many times the program is run. This is due to two factors, the first being that preservation of the amino acid code only allows a limited number of mutations, and thus there may not be a mutation that can silence all SR proteins in a particular motif. The program will try all possible combinations of nucleotides, but if it is not possible to mutate, it will output the particular motif with a message indicating that there is no combination of nucleotides that can silence an ESE for a particular SR protein at a particular motif. The second factor which may lead to ESEs in the output of the program is related to the phenomenon discussed above, whereby a silenced ESE for a particular SR protein in one motif may introduce another ESE for a different protein in the surrounding area. In this case, the program will try up to 10 times to mutate: if not all ESEs for all SR are silenced, then the program will choose (at random) one of the ESEs to retain in the output. Running the DNA sequence through the program can allow this ESE to be silenced, but will result in the second ESE appearing, as only one can be silenced at the expense of another.

An additional consideration is that The EXONerator's scoring for ESEs, the threshold levels and the motif strengths were all based on the web-based program, ESEfinder [1]. Therefore, all of their limitations also apply here. In particular is the limitation concerning the limited identification of ESEs for only four human SR proteins (SF2/ASF, SC35, SRp40 and SRp55), as detailed analysis of ESE motifs and strength of binding of motifs for other SR proteins is still not available. However, there is currently much work dedicated to identifying motifs for other RNA binding proteins that can predict splicing. For example, a recent paper has identified splicing factor binding sites for 15 other RNA binding proteins (in addition to the four currently considered in The EXONerator) [119]. As more information becomes available for the strength of binding of these (and other) proteins, this data can be incorporated easily into The Exonerator by adding the motif "scores" of each set of nucleotides for additional proteins in the form

of more arrays in the PERL code. Thus, the functionality of The Exonerator in predicting ESE sequences can be continually expanded as more information becomes available.

However, utilization of the detailed information currently available in the ESEfinder, combined with The EXONerator, to produce a sequence with silenced ESEs that can then be tested in the lab is an extremely valuable tool to any investigator who wishes to explore the location and/or functional significance of an ESE in any coding sequence of interest. As more information about SR proteins surfaces, the PERL coding can be easily modified to include information about other RNA binding proteins as it become available.

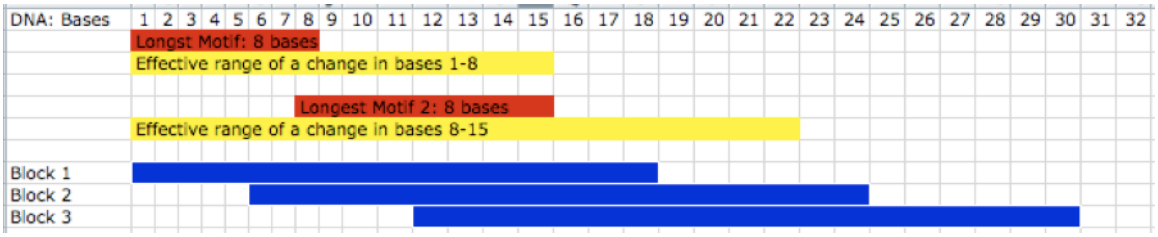


Figure 4.1: Illustrating the logic of dividing DNA string into blocks in The EXONerator. Two example motif positions (red) are exemplified, with numbers indicating the position of a DNA base. As is evident in the figure, a single base change in a motif of 8 bases long (the longest motif for the SR proteins considered) can affect 7 bases upstream and 7 bases downstream (illustrated in yellow); thus an entire block of bases must be taken into account when mutating DNA. The most efficient way to break down the DNA sequence was found by way of trial and error: the DNA string is divided up in blocks of 18 bases, with a shift of 9 bases downstream each time so that each preceding block overlapping with the next block with the last 12 bases (illustrated in blue). Once the first block is changed, the modified block replaces the original, so that the second block includes any modifications made, and so on.

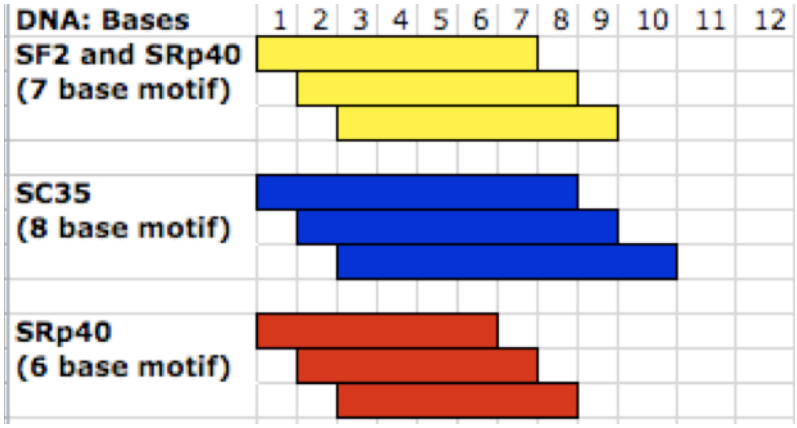


Figure 4.2: Illustrating the shift in bases as motif scores are checked. Motifs for each protein (colored blocks) are checked at every position in the 18 base pair block, starting at position one. Once the first score is calculated and changes are made, the program then shifts one block to the right and calculates scores/makes changes if necessary for the second position, and so on. Once the program gets to a position where there are too little bases to make up a motif, the program ignores the remaining bases (for example, SF2 has a 7 base motif, so when there are six bases left in a block, the program ceases to check SF2). These unchecked bases are re-checked within the next block, since they overlap, so it does not present a problem of overlooked ESE sequences.

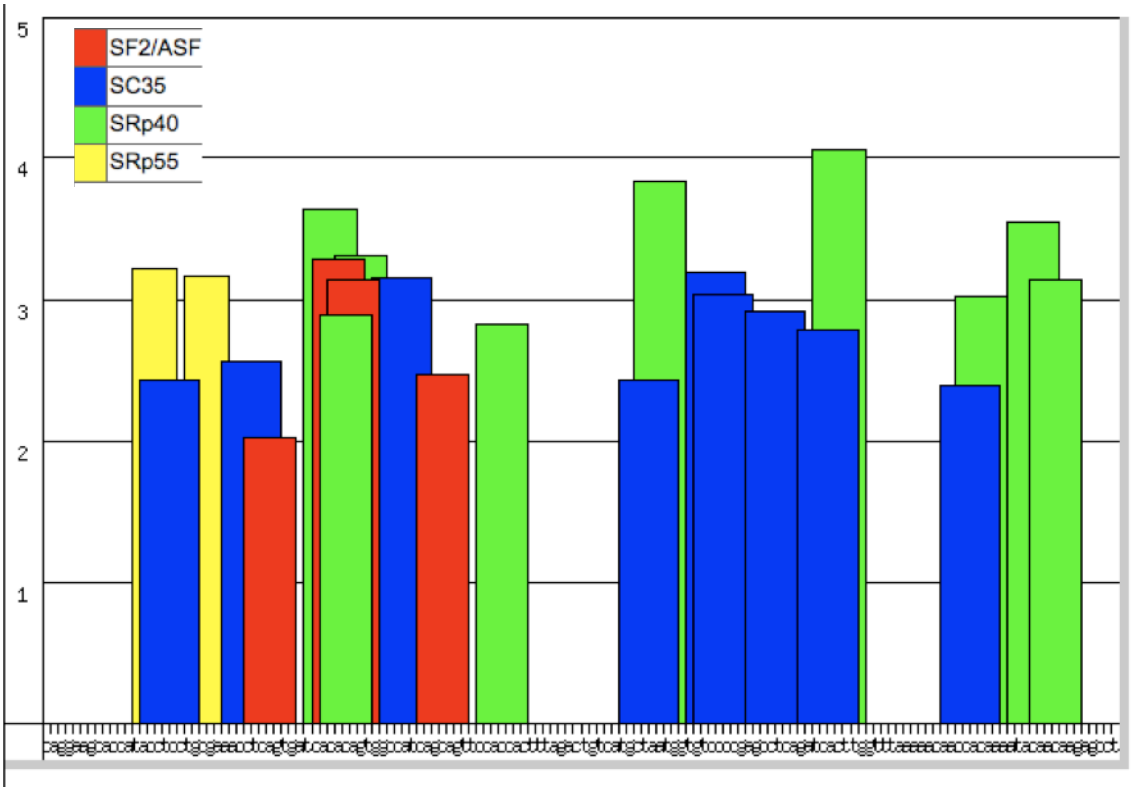


Figure 4.3: Predicted ESEs in the 144 base pair long coding region of Flt1 exon 14. ESEs illustrated by bars on x axis (DNA sequence), with width of bars indicating motif length for each SR protein. The y axis indicates score of predicted ESE, with higher score indicating higher strength of predicted binding of SR protein. As is evident in the figure, there are 24 predicted ESEs in Flt1 exon 14 coding region (Illustrated with ESEfinder program [1]).

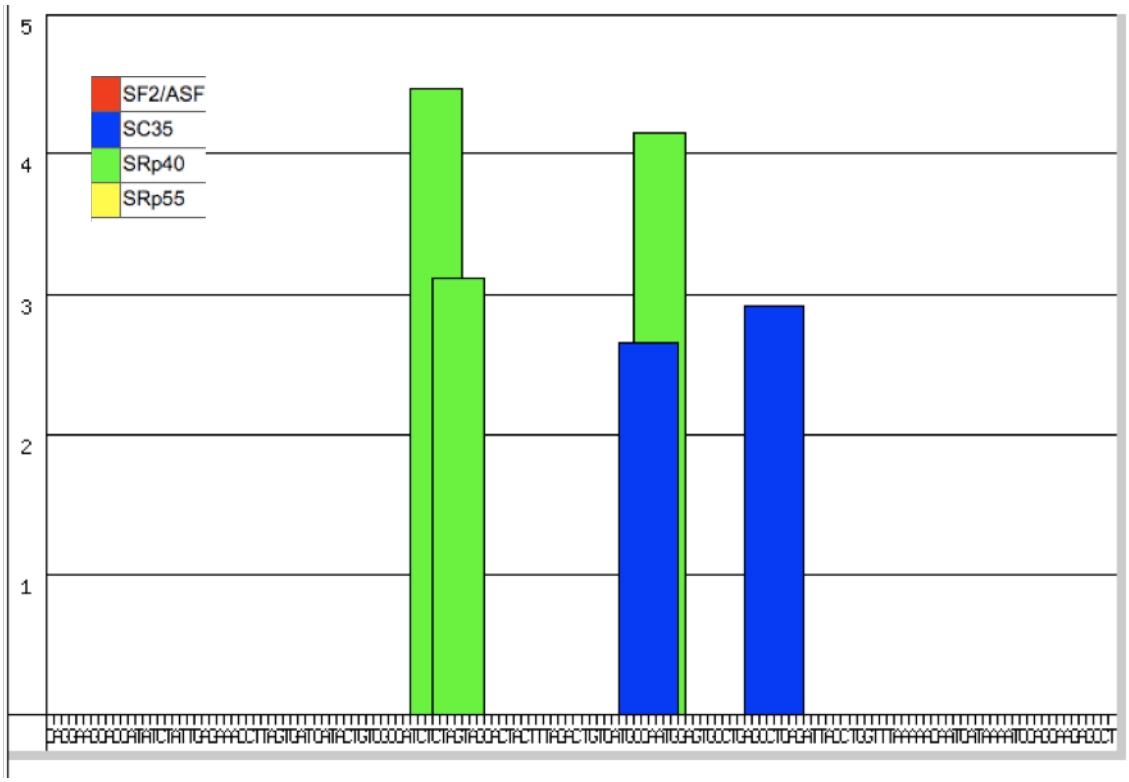


Figure 4.4: First generated sequence of The EXONerator (illustrated by the ESEfinder program [1]) After running the 144 base-pair long coding region of exon 14 (Figure 4.3, above) through The EXONerator, only 5 predicted ESEs remain from the initial 24 predicted ESEs. (Predicted ESEs illustrated by bars on x axis (DNA sequence), with width of bars indicating motif length for each SR protein. The y axis indicates score of predicted ESE, with higher score indicating higher strength of predicted binding of SR protein).

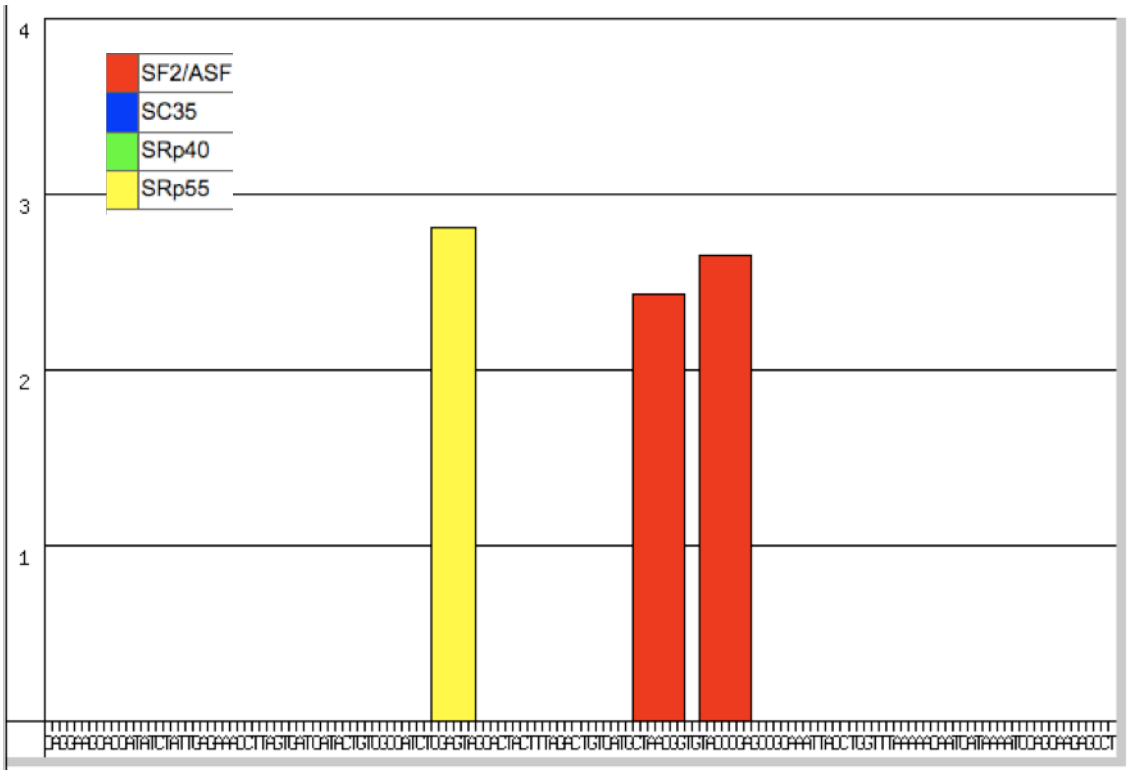


Figure 4.5: Second generated sequence of The EXONerator (illustrated by the ESEfinder program [1]). After running the changes made to the coding region of Flt1 exon 14 (in the first output, see Figure 4.4,), all 5 predicted ESEs were silenced, but three different ones were introduced. (Predicted ESEs illustrated by bars on x axis (DNA sequence), with width of bars indicating motif length for each SR protein. The y axis indicates score of predicted ESE, with higher score indicating higher strength of predicted binding of SR protein).

No ESE found in the sequence.

Figure 4.6: Final generated sequence of the EXONerator (illustrated by the ESEfinder program [1]). After running the changes made in the second output (Figure 4.5,), all ESEs in the Flt1 coding region of exon 14

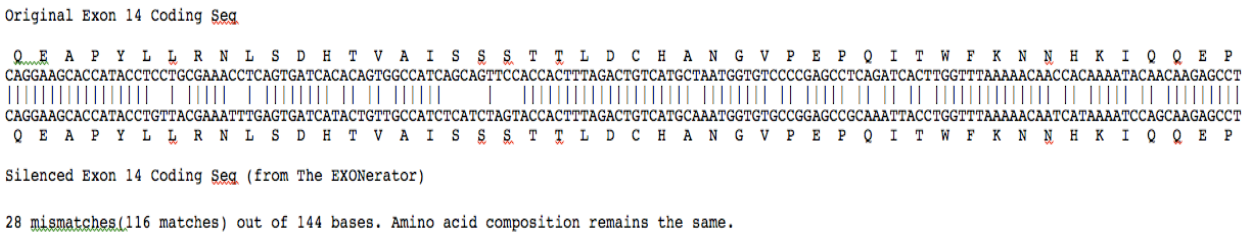


Figure 4.7: Comparison of Exon 14 144bp coding sequence (top) with final generated sequence of The EXONerator (bottom) in which all ESEs were silenced. As is evident in the figure, there were 28 changes in the base pair sequence, with no change in the amino acid sequence.

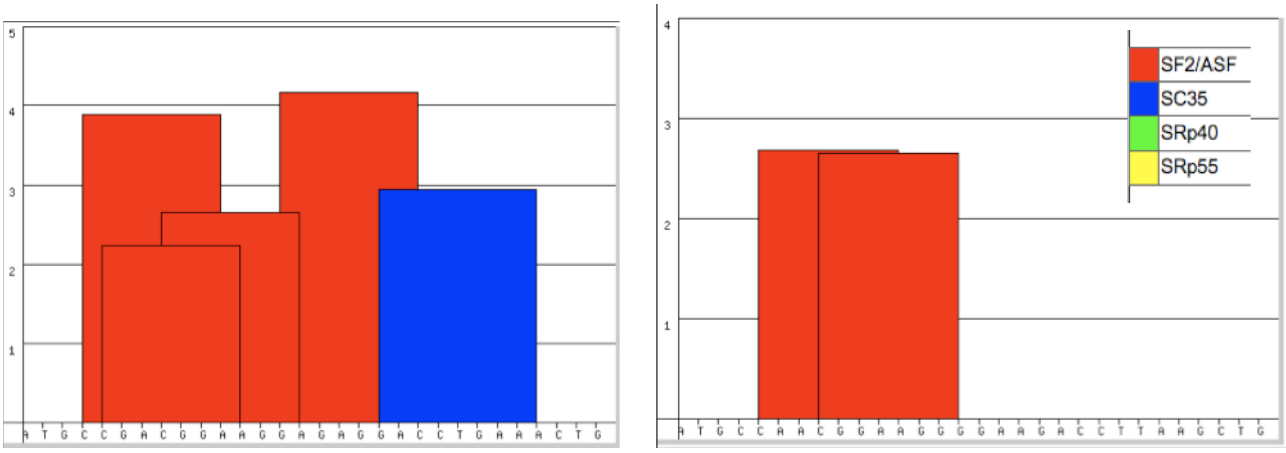


Figure 4.8: Comparing Manually Silenced ESEs with EXONerator silenced ESEs in Exon 13 region (illustrated by the ESEfinder program [1]). Figure on the left shows five predicted ESE sequences in exon 13 region. Manual changes made to silence ESEs silence all but two of the ESE sequences. Running the same unmodified sequence through the ESE finder resulted in all five ESEs being silenced. (Predicted ESEs illustrated by bars on x axis (DNA sequence), with width of bars indicating motif length for each SR protein. The y axis indicates score of predicted ESE, with higher score indicating higher strength of predicted binding of SR protein).

```
Unmodified Ex 13:          ATGCCGACGGAAGGAGAGGACCTGAAACTG
Manual Modification:       ATGCCAACGGAAGGGGAAGACCTTAAGCTG
EXONerator Modification:   ATGCCGATGGTAGGCGGAGACCTTAAACTG
```

Figure 4.9: Comparing Sequence for Manually Silenced ESEs with EXONerator silenced ESEs in Exon 13 region. The wild type sequence is shown in black, with changes in red. As is evident in figure, the manual modification resulted in 5 base changes, with two ESEs not silenced (Figure 4.8) while The EXONerator changed 8 bases, resulting in silencing all ESEs in the sequence.

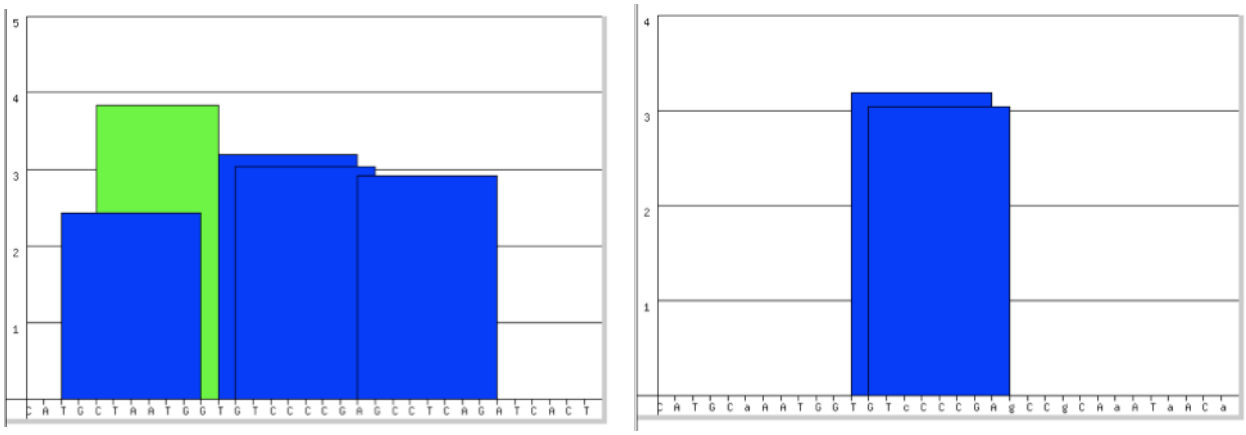


Figure 4.10: Comparing Manually Silenced ESEs with EXONerator silenced ESEs in Exon 14 region (illustrated by the ESEfinder program [1]). Figure on the left shows five predicted ESE sequences in exon 14 region. Manual changes made to silence ESEs silence all but two of the ESE sequences. Running the same unmodified sequence through the ESE finder resulted in all five ESEs being silenced. (Predicted ESEs illustrated by bars on x axis (DNA sequence), with width of bars indicating motif length for each SR protein. The y axis indicates score of predicted ESE, with higher score indicating higher strength of predicted binding of SR protein).

```
Unmodified Ex 14:          GCTAATGGTGTCCCCGAGCCTCAGATCACT
Manual Modification:       GCAAATGGTGTCCCCGAGCCGCAAATAACA
EXONerator Modification:   GCAAATGGTGTGCCTGAGCCGCAAATAACA
```

Figure 4.11: Comparing Sequence for Manually Silenced ESEs with EXONerator silenced ESEs in Exon 14 region. The wild type sequence is shown in black, with changes in red. As is evident in figure, the manual modification resulted in 5 base changes, with two ESEs not silenced (Figure 4.8) while The EXONerator changed 6 bases, resulting in silencing all ESEs in the sequence.

Chapter 5: Multi-Species Comparison of the Flt1 gene

Abstract

There are many evolutionary forces at work that can contribute to genetic diversity, including mutation, genetic drift, gene flow and natural selection. Exons tend to be highly conserved between species, as changes in coding sequence can disrupt function. However, when high conservation among species is observed in non-coding regions, this may imply a regulatory function. Alignment of sequences from different species allows identification of highly conserved sequences, which can help identify potential regulatory elements within noncoding regions. In this study, we studied an important regulatory region in the (exon 13 – intron 13 – exon 14) Flt1 gene in order to identify domains that may play a role in the alternative processing to produce full length Flt1 or sFlt1. Sequences were collected from the databases for all available species for which an equivalent Flt1 genomic sequence was found, aligned to the human sequence using MLAGAN and visualized with VISTA browser. In addition, a rankVISTA analysis was performed to identify the statistically significant conserved regions. Several regions of significant sequence correspondence were found in the three coding regions (the two exons and coding region of intron 13). There were also several regions within intron 13 that showed very high sequence identity, two of which correspond to regions previously identified in the mouse genome as important regulatory regions containing cleavage-polyadenylation signals. These regions of high identity within Intron 13 have likely been conserved evolutionarily due to a regulatory function in the pre-mRNA processing involved in Ftl1/sFlt1 expression. Further analysis of conserved regions in intron 13 will allow determination of whether conservation occurred as a result of a random process or due to evolutionary forces working to conserve essential domains that can influence pre-mRNA processing and affect the levels of Flt1 to sFlt1 expression.

Introduction

Sources of Genetic Diversity

Genetic diversity is the variation that occurs at the level of individual genes. There are many evolutionary forces at work to produce persistent genetic differences, both within the same species as well as between different species, but there are four main processes: mutation, genetic drift, gene flow (also referred to as genetic exchange) and natural selection [120].

The process of DNA replication is not error-free, and mutations often arise. The error rate of DNA polymerases is approximately one error per 10^4–10^5 nucleotides, with most errors being corrected by 3′→5′ exonucleolytic proofreading activity and post-replication mismatch repair, but uncorrected errors can lead to mutations [121, 122]. These random mutations are the basis for evolution as they may give rise to functional diversity [120, 121]. Mutations accumulate gradually over time, so that the longer the time that has elapsed since the species diverged, the more mutations (and thus genetic variation) that will be apparent [121].

Genetic drift, a change in allele frequencies of a population can occur via the founder effect (genetic make-up of a new population is established by only a few founding members, and thus may differ greatly from the genetic make-up of the original population) or genetic bottleneck (the original population suffers a drastic loss in the number of individuals, and thus a new population arises from the few remaining survivors) [123]. Thus, by chance, some alleles can have their frequency changed or even be lost from the population [120].

Gene flow (or genetic exchange) is the introduction of new genetic material into a population, which can also introduce mutations that originated elsewhere [120]. Addition of new genetic variants to an established gene pool serves to increase the genetic diversity [120].

Natural selection is a very strong evolutionary force that occurs as a result of adaptation (hereditary changes in the genotype that results in increased phenotypical fitness to the environment) [121, 124]. This serves to standardize the frequency of

alleles within a specific population but different environments will result in different alleles being chosen for natural selection [121].

Comparative Genomics

In addition to elucidating evolutionary relationships, recognizing conserved regions through the alignment of sequences from different organisms can be a valuable tool in identifying coding regions, as well as regulatory elements within noncoding regions [125]. Natural selection can be positive or negative for a specific allele. Negative selection serves to remove mutations that may reduce fitness of the organism or those that may even be fatal, while positive selection will work to preserve the essential regions of the genome, as well as those that give the organism a survival advantage. As a result, most exons between species are highly conserved since changes in the exons can result in changes in protein sequence, which in turn can result in the disruption of function [124]. On the other hand, noncoding regions such as introns accumulate many changes in base pair sequence, and can vary dramatically between species [124]. The current availability of databases with genomes sequenced from many different species (such as NCBI Entrez Genome Project [71] and the Human Genome Sequencing Center at Baylor University [72]), combined with the availability of multi-species alignment software tools have allowed for detailed species comparison studies. Such studies can be very useful in identifying the importance of regions in a gene that are essential to function (especially noncoding regions, such as introns) and also in identifying SNPs within a gene that have the potential to affect the function of a gene [59]. The identification of such evolutionarily conserved regions allows for further studies to identify their function in gene and protein function and, in the case of the Flt1 gene, also allow for identification of domains that may play a role in the alternative processing to produce full length Flt1 or sFlt1 [126].

It is important to note that comparative genome studies can indicate important regulatory regions, as conservation of certain domains points to functional importance [125]. However, since the processes that lead to variation (for example, mutations) are random, some regions may be conserved and yet have little or no functional significance [125]. After identifying these conserved regions in the Flt1 gene, further

analysis will allow determination of whether conservation occurred as a result of a random process or due to evolutionary forces working to conserve an essential domain.

Materials and Methods

The process of identifying conserved regions involves three steps. The first step is identifying the region one wishes to study and collecting sequences for the species under investigation. Then, these sequences need to be aligned using appropriate software, depending on the length of the region, the number of sequences and how closely related the species under study are. The last step, which is helpful but not essential, is visualizing the alignment, so that one can identify easily the conserved regions

Sequence Retrieval

Using the same sequence described in Chapter 2 (Exon13- Intron 13-Exon-14 of the Flt1 gene, with portions of Intron12 and Intron14), sequences were collected from the databases for all available species for which an equivalent Flt1 genomic sequence was found (Table 5.1). For other species, such as the sheep (*Ovis aries),* frog *(Xenopus borealis*), opossum (*Monodelphis domestica*) and platypus (*Ornithorhynchus anatinus*), only an mRNA sequence for the Flt1 gene was available, so it could not be used in analysis as there was no intron sequences for comparing intron 13.

Other sequences were identified by NCBI Entrez to be equivalent for the Flt1 gene in organisms for which there is no circulatory system, including the btl gene in the fruit fly (*Drosophila melanogaster*, sequence NM_168577 from NCBI Entrez), the egl-15 gene in the roundworm (*Caenorhabditis elegans*, sequence NM_077441 from NCBI Entrez), and the BCK gene in yeast (*Saccharomyces cerevisia*, sequence NC_001142 from NCBI Entrez). In addition, the Flt1-equivalent sequence for the Zebrafish (*Danio rerio*) was identified (NW_634345 from NCBI Entrez), but very little is known about it's function, and there is almost no sequence similarity with human Flt1, and it was therefore excluded from the analysis.

There is no information available for the BCK gene in *S. cerevisia*. In *D.*

melanogaster, the btl gene codes for the Fibroblast growth factor (FGF) receptor (FGFR), which controls the migration of glial, mesodermal, and tracheal cells [127]. In *C. elegans*, the egl-15 gene produces two forms of the protein via alternative splicing: EGL-15(5A) (required for the gonadal chemoattraction of the migrating sex myoblasts) and EGL-15(5B) (essential for viability) [128]. This suggests that the function of the Flt1 gene has evolved along with its sequence.

Aligning the Sequences and Visualizing Alignment

Once the sequences had been retrieved, identifying an appropriate software tool to align sequences properly and identify conserved regions in the Exon 13-Intron 13-Exon 14 region of Flt1 region was the next step. There are many commercially available software tools that can be used for this purpose, but the VISTA browser using the MLAGAN alignment program was identified as a very powerful tool for this type of analysis.

The VISTA program was chosen for several reasons: it allows simultaneous visualization of several alignments – up to several megabases - on the same scale, it uses a continuous curve to indicate the level of conservation and it is very effective at handling gaps [129].

The alignment was done using MLAGAN, a program that works with the VISTA browser. This alignment method was chosen because, unlike most programs currently available, it boasted two very important features essential to this analysis: it is highly efficient in multiple sequence alignment and these sequences can be as extremely distantly related evolutionarily as fungi and human [130]. MLAGAN is based on the pair-wise alignment program, LAGAN. LAGAN aligns two sequences by creating a local alignment, constructing a rough global map by chaining a subset of the local alignments, and then computes the final alignment based on the best alignment around the rough global map. MLAGAN aligns sequences via intermediate multiple alignments with LAGAN, before removing each sequence from the multiple alignment sequentially, and realigning until it reaches a point where no significant improvements can be made [130].

Results

Aligning the genomic DNA sequences for the region of interest of the Flt1 gene revealed several regions of significant sequence correspondence (Figure 5.1). As expected, Exon 13, the 5' coding region of Intron 13 and Exon 14 showed very high identity (Table 5.2). Interestingly, the coding portion of Intron 13 is very highly conserved; its sequence shows higher sequence identity than both surrounding exons.

Surprisingly, there were several other portions of Intron 13 that showed very high identity in spite of being noncoding, indicating regions of potential functional significance in the gene (Table 5.3). A portion of intron 13 that was not conserved was chosen at random to show how significant the difference in identity between the conserved regions and the nonconserved regions.

In addition to aligning the sequences with the VISTA browser, a rankVISTA alignment was made (Figure 5.2). rankVISTA estimates the evolutionary rate of change between multiple sequences in neutral, noncoding regions and compares it to other regions to find ones with a slower rate of change, which it then identifies as a conserved region [80]. The output of rankVISTA is a bar graph, whereby the height indicates the statistical significance of the region identified as conserved [$-\log^{10}$(P-value)]. Thus, a height of 5 indicates that a probability of less than 10^{-5} for a randomly-evolving 10-kb segment to show this level of conservation by chance [80]. As is evident from the figure, rankVISTA identified 5 regions as statistically significant conserved regions: Exon 13 combined with almost 200 bases of 5' Intron 13, a 94 base region of Intron 13 (bases to 578-672) with a score of 3, a middle region (bases 1386-1746) of Intron 13 with a score of 5, a 132 base region in 3' Intron 13 (4011-4143) with a score of 1.5 and the 3'end of Exon 14, with a score of 1.

Discussion

Examining conserved regions in genes can provide insight into importance of specific domains of the gene. This is particularly true for introns, as the lack of a coding function has left much room for evolutionary change. However, the 5' region of Intron 13

of the Flt1 gene appears to have exon-like functionality in the case of sFlt1 expression, and the observed conservation of this region across species is expected to be consonant with exon conservation. In fact, the coding sequence of Intron 13 shows equal or higher degrees of conservation than both surrounding exons in every species.

There is almost no information for expression of the Flt1 gene in most species (apart from human Flt1 expression), the exceptions mainly being mouse and rat, and (to a much lesser extent), pigs, cows and dogs. In mice, Flt-1 is essential for growth and development of embryos, as inactivation of a single VEGF allele results in embryonic lethality between day 11 and day 12 [14, 15]. In addition, both mouse and rat Flt1 have been shown to act in a similar manner to the human Flt1 gene, as mouse/rat Flt1 induces endothelial cell division and morphogenesis, and modulates signaling through the Flk1 receptor (the mouse/rat equivalent of KDR) and branching of blood vessels [131-135]. In pigs, cows and dogs, there are very few studies on Flt1 expression, but the ones available suggest a very similar expression profile to human Flt1 I, as well as similar functions in physiological angiogenesis, as well as pathological [136-138]

Previous work in the laboratory has identified 85.6% identity between the mouse and human sequence in the region of identified cleavage-polyadenylation signals, which is much higher than the identity observed in other non-coding regions of intron 13 [2]. This region is equivalent to the “3’ Intron 13 region” (Table 5.1), a portion of which was scored by rankVISTA as a region of significant conservation (score = 1.5). In addition, the “central” region of Intron 13 (Table 5.1), also identified by rankVISTA as a region of significant conservation (score = 5), corresponds to a region identified previously in the mouse genome as having two cleavage-polyadenylation signals (#1/2) [2]. In previous studies in the lab, it was suggested that these cleavage-polyadenylation signals were not major ones: when deleted, they produced little effect (suggesting that other signals can compensate), and when deleted alongside the more important signals #5 and #6 (which resulted in a 77% decrease of sFlt1), there was no additional decrease in expression of sFlt1 levels [2]. However, their high levels of conservation, especially when compared to the conservation of already established essential signals #5 and #6, could suggest several things. Perhaps the use of these signals may be essential in a

certain developmental stage, as they may be preferentially used at some point in development. Alternatively, their use may be tissue specific. Another inference from this extremely high conservation may point to their preferential use in different species other than mouse, as it may be that different species preferentially utilize different cleavage-polyadenylation sequences. This is highly likely in the chicken, as it is the only organism that does not show conservation at signals #5 and #6, suggesting that they depend on other cleavage-polyadenylation sites. Since the chicken genome does show high conservation and signals #1 and #2, these may be the signals that are used in place of #5 and #6.

Identification of functional significance in the mouse, as well as a high degree of conservation among species in these regions, points to high functional significance of this region of intron 13 among all species. These regions of high identity within Intron 13 have likely been conserved evolutionarily due to a regulatory function in the pre-mRNA processing involved in Ftl1/sFlt1 expression. Now that these regions have been identified with a very high indication of their likely significant roles, they can be tested experimentally to determine their influence on pre-mRNA processing and, consequently, how they affect the levels of Flt1 to sFlt1 expression.

Table 5.1: Species for which the Flt1 sequence of interest (Exon13- Intron 13-Exon-14) was collected from the available databases.

Species	Sequence Name	Source
Human (*Homo sapiens*)	NT_024524.13	NCBI Entrez
Chimpanzee (*Pan troglodytes*)	NW_115447	NCBI Entrez
Rhesus Monkey (*Macaca mulatta*)	Mmul0.1Contig62252	Baylor Database
Horse (*Equus caballus*)	NW_001867382.1	NCBI Entrez
Dog (*Canis familiaris*)	NW_139897.1	NCBI Entrez
Cow (*Bos taurus*)	NW_001493038.1	NCBI Entrez
Mouse (*Mus musculus*)	NT_039324	NCBI Entrez
Rat (*Rattus norvegicus*)	NW_047366.2	NCBI Entrez
Chicken (*Gallus gallus*)	NW_060243	NCBI Entrez

Table 5.2: Percentage of conservation of coding sequences across species, as compared to the human sequence (from the VISTA browser, aligned with MLAGAN).

Species	Coding Sequence Identity to Human Sequence		
	Exon 13	Coding Region of Intron 13 (nt1-99)	Exon 14
Chimpanzee (*Pan troglodytes*)	99.4%	100%	100%
Rhesus Monkey (*Macaca mulatta*)	98.4%	100%	98.0%
Horse (*Equus caballus*)	92.6%	97.0%	91.8%
Dog (*Canis familiaris*)	90.9%	96.0%	89.0%
Cow (*Bos taurus*)	79.9%	92.0%	90.5%
Mouse (*Mus musculus*)	80.9%	91.0%	82.1%
Rat (*Rattus norvegicus*)	80.6%	97.0%	82.8%
Chicken (*Gallus gallus*)	73.5%	92.0%	72.0%

Table 5.3: Percentage of conservation of species sequences, as compared to the human sequence, for noncoding regions of Intron 13 (from the VISTA browser, aligned with MLAGAN). Numbers refer to base numbers in Intron 13. NC = “nonconserved”.

Species	Intron 13 Identity to Human Sequence			
	5’ region (1-890)	Central Region (1386-1746)	3’ region (3875-4316)	Random non-coding (1894-2212)
Chimpanzee (*Pan troglodytes*)	98.5%	99.4%	99.1%	98.4%
Rhesus Monkey (*Macaca mulatta*)	98.3%	98.6%	99.8%	92.5%
Horse (*Equus caballus*)	91.3%	94.1%	93.5%	78.2%
Dog (*Canis familiaris*)	85.0%	94.7%	90.2%	71.5%
Cow (*Bos taurus*)	82.5%	89.3%	84.5%	76.8%
Mouse (*Mus musculus*)	75.1%	73.3%	84.9%	0.0%
Rat (*Rattus norvegicus*)	71.3%	76.1%	80.6%	0.0%
Chicken (*Gallus gallus*)	75.2%	79.9%	0.0%	0.0%

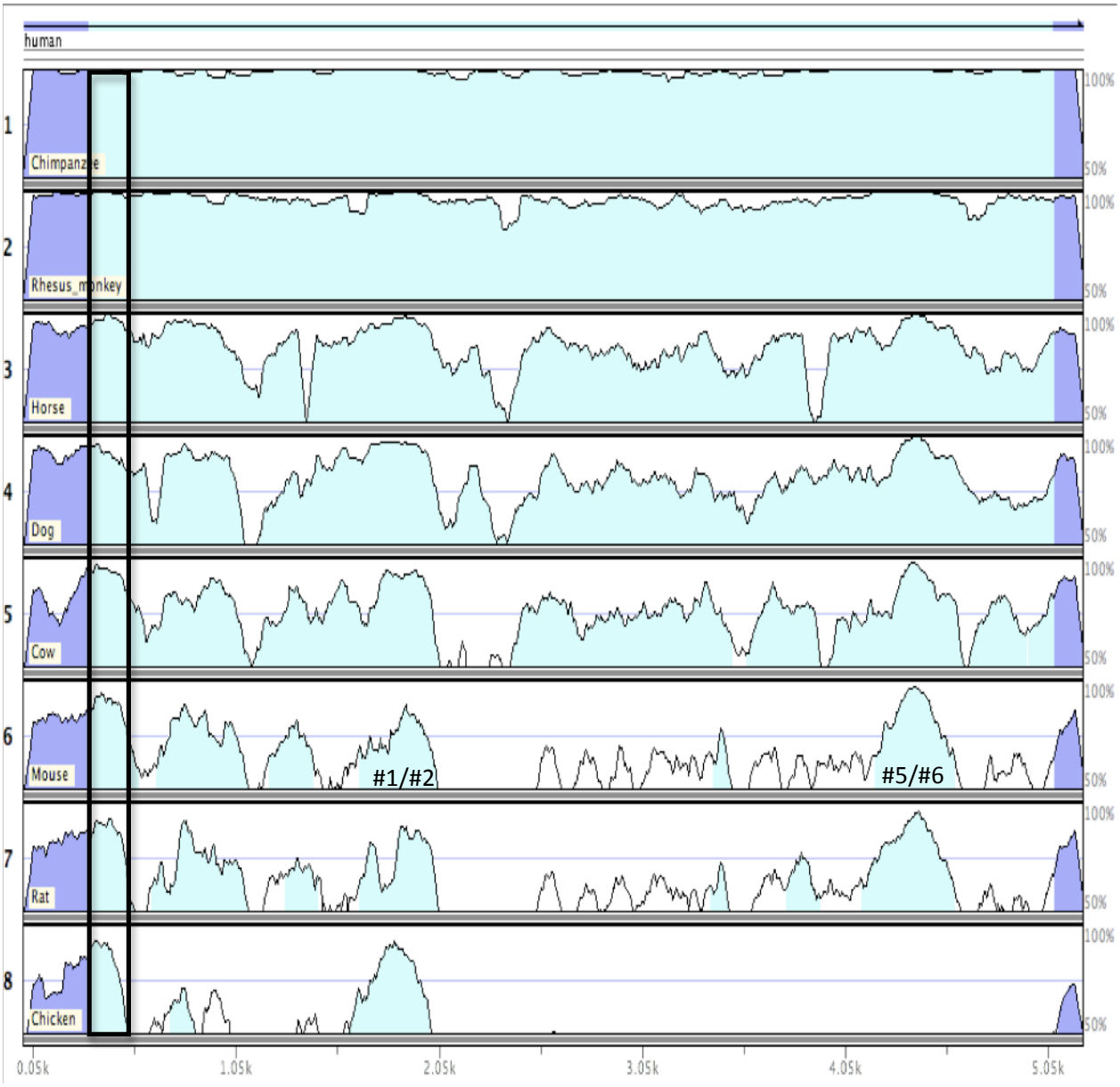


Figure 5.1: Plot of multiple aligned sequences of the Flt1 Exon13 – Exon 14 region with MLAGAN and VISTA Browser. Exons are colored dark blue, while introns are light blue. Height of curve indicates degree of identity to human sequence. Curves with no color indicate some identity, but not enough to be significant. Black bar indicates coding region of intron 13. As is evident in the figure, Exon 13, the 5' coding region of Intron 13 and Exon 14 show very high identity, and several regions in intron 13 also show high identity in spite of these regions being noncoding. The two highest conserved regions have already been previously identified as being important cleavage-polyadenylation signals in the mouse ([2], annotated on mouse genome). Other regions of high conservation may point to functional significance.

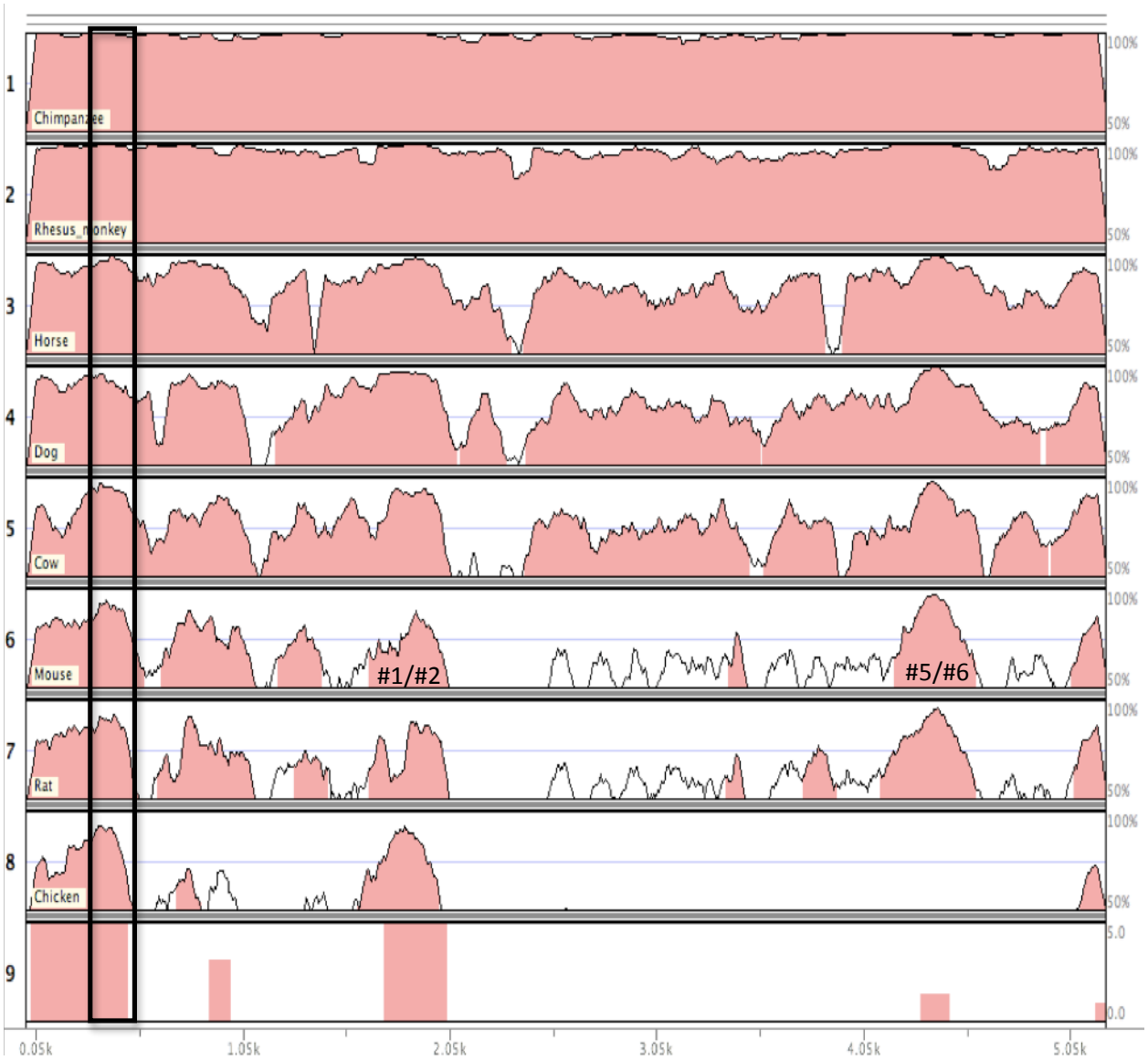


Figure 5.2: Plot of multiple aligned sequences of the Flt1 Exon13 – Exon 14 region with MLAGAN, rankVISTA and the VISTA browser. A block in Row 9 indicates regions identified as significantly conserved by rankVISTA: the higher the block, the higher the statistical significance of conservation. As is evident, rankVISTA identified 5 regions as statistically significant conserved regions: Exon 13 combined with almost 200 bases of 5' Intron 13, a 94 base region of Intron 13 (bases to 578-672) with a score of 3, a middle region (bases 1386-1746) of Intron 13 with a score of 5, a 132 base region in 3' Intron 13 (4011-4143) with a score of 1.5 and the 3'end of Exon 14, with a score of 1. Two of the ranked regions in intron 13 have already been previously identified as being important cleavage-polyadenylation signals in the mouse ([2], annotated on mouse genome).

Chapter 6: General Discussion

Synopsis of Results

Understanding how angiogenesis is regulated is an important step in implementing new therapies for diseases related to aberrant angiogenesis, such as cancer. Due to the importance of VEGF activity (and the VEGF receptors through which it mediates its biological response), blocking its effect has become a main focus of such therapies. Overexpression of sFlt1 inhibits tumor growth, and thus therapies that manipulate sFlt1 expression may be an important therapy approach. In order to do so, one must understand how Flt1 vs. sFlt1 expression is regulated. In this study, we have attempted to understand the effects of naturally occurring polymorphisms in the regulation of Flt1 and sFlt1 expression.

Querying the available information in the databases for SNPs in the Flt1 region resulted in the identification of only a few SNPs. We tested the functional effects of the exonic SNPs found by constructing a human Flt1 minigene construct and introducing these SNPs into the minigene construct. In addition, we tested the effects of silencing the ESE sequences surrounding these SNPs.

SNPs lead to variation between individuals in disease susceptibility, severity and progression, as well as response to therapies [87]. Characterizing SNPs in genes that contribute to a disease can be extremely helpful. Individuals with SNPs that have an effect on disease susceptibility can be aided in prevention (if possible) and monitored for early detection of disease. Doctors may be able to tailor therapies on a case-by-case basis for individuals with SNPs that affect severity and progression of a disease, resulting in a higher efficiency of treatment. This is also important in if SNPs have an effect on the efficacy and/or side effect of certain drugs used to treat the disease, as ineffective drugs can be avoided and substituted with other drugs that may be more effective.

However, there is very little study into the SNPs in the Flt1 gene, and thus no information on their functional effects in individuals. Our results indicated that the studied SNPs (in the context of the minigene system) lie in regulatory regions that affect

the relative rates of RNA processing for Flt1 and sFlt1, which can have many implications for therapy if this trend is also found in native context.

Studying ESEs is somewhat difficult, as mutating the ESE region (while preserving the amino acid sequences) is a time consuming process. To this end, we have created The EXONerator, a PERL program that automates this process: simply enter a coding sequence of interest and receive a generated sequence with most (if not all) ESEs silenced, but still preserving the amino acid sequence. Sequences with silenced ESEs can then be tested in the lab for functional effects. This ease of producing silenced ESE sequences will be a valuable tool in further studying the regulatory regions in the exons of Flt1 that have an effect on Flt1 and sFlt1 mRNA processing, as one can identify which ESEs are functional and which predicted ESEs are theoretical. In addition, this is a valuable tool that any investigator can use to apply to their coding sequence of interest.

Another aspect of this study was to expand the SNP information available in the Flt1 gene. An independent SNP screen of the Flt1 exons 13 and 14 was undertaken to confirm the non-validated SNPs reported by the SNP databases, as well as attempt to identify new SNPs in these exons. We were able to confirm several existing SNPs, and identify some potential unreported SNPs, but this is only a start. A more thorough screen of SNPs in the Flt1 region, especially in previously identified [2] regulatory regions in intron 13, is essential. This is doubly important, given the functional effects of SNPs that were shown in our minigene model: it is likely that SNPs in the intron 13 region also play a role in regulating Flt1 and sFlt1 mRNA levels.

The final area of study involved examining conserved regions in the Flt1 region of interest to provide insight into importance of specific domains. This was especially illuminating for sequence of intron 13. I expected to find high conservation in the 5' intron 13 coding region of intron 13, as it has exon-like functionality, and indeed its conservation across species is higher or comparable to the conservation of the surrounding exons. In addition, we identified several regions of extremely high conservation across species in the intron, which coincided with regions identified previously in the lab as important cleavage-polyadenylation signals in the mouse genome[2]. Identification of functional significance in the mouse, as well as a high

degree of conservation among species in these regions, points to high functional significance of this region of intron 13 among all species.

In conclusion, the results of this study support our initial hypothesis: naturally occurring polymorphisms in key regulatory regions of the Flt1 locus can alter the relative expression of Ftl1 and sFlt1 mRNA.

Future Work

Confirming and extending current SNP data is an important step in future Flt1 studies. Due to the scarce information available, one is unable to identify which SNPs may have a potential functional effect (which can then be tested) and which SNPs may be misreported. Thus, a more thorough screen of the Flt1 exon 13-intron 13-exon 14 region would likely answer many of those concerns. High-throughput next generation sequencing technologies, such as the ABI SoliD and Illumina, would prove essential in such a task.

Illumina Sequencing technology utilizes a novel and highly accurate approach to sequencing, whereby randomly fragmented DNA is sequenced in parallel - millions of fragments at a time - using DNA sequencing-by-synthesis technology [139]. Sequences are then aligned against a reference genome, and differences can be detected using specialized software. The SOLiD method utilizes sequencing by ligation with random fluorescently labeled probes whereby every base is read in two different dinucleotide frames [140]. Sequences are also analyzed by specialized software for the SOliD platform. Sequencing accuracy is very high for both platforms (less than 1% error rate), and cost per sample continues to drop, which may make this an affordable option.

Newly identified SNPs, as well as those currently reported in the databases (all of which are still not validated) can be tested for functional effects in human Flt1 minigene models such as the one described in this study. This minigene model can also be used to study the functional effects of regions of high conservation among species. By randomly mutating regions of high conservation, the functional significance of these domains could be studied. In addition, using interspecies minigenes would be an interesting avenue to explore. For examples, one could replace the human sequence of

a portion of the intron with that of a closely related species and that of a distant one and test the effects of such a replacement.

Besides testing functional effects in minigene models, one could also extend the study and test functional effects in the entire Flt1 gene, in an endothelial cell model. This would be attempted by purifying endothelial cells from an adipose tissue biopsy taken from an individual with a non WT SNP variant, as well as from an individual with a WT SNP variant, and commencing with analyzing expressing levels of these purified endothelial cells to compare expression levels in each case. Such a study would give a much clearer idea of the effects of a SNP on native Flt1 and sFlt1 levels, as opposed to an extrapolation of behavior such as that in our minigene model.

The use of The EXONerator on the Flt1 gene sequence would also be an interesting study to pursue. As shown in Chapter 4, all the ESEs in exon 14 were successfully silenced. If this silenced sequence was tested in a minigene context, Flt1 levels would be expected to drop dramatically, as signals for splicing decrease, and cleavage-polyadenylation would likely increase resulting in a reciprocal increase of sFlt1 levels.

Understanding the regulation of mechanisms involved in Flt1 or sFlt1 expression is an important step towards manipulating their levels, which would allow some control over rates of angiogenesis. This would, in turn, allow for more effective therapies in diseases where aberrant angiogenesis plays a crucial role.

References

1. Cartegni, L., et al., *ESEfinder: A web resource to identify exonic splicing enhancers.* Nucleic Acids Res, 2003. **31**(13): p. 3568-71.
2. Huckle, W.R. and R.I. Roche, *Post-transcriptional control of expression of sFlt-1, an endogenous inhibitor of vascular endothelial growth factor.* J Cell Biochem, 2004. **93**(1): p. 120-32.
3. SNP, N.; Available from: http://www.ncbi.nlm.nih.gov/snp.
4. Blencowe, B.J., *Exonic splicing enhancers: mechanism of action, diversity and role in human genetic diseases.* Trends Biochem Sci, 2000. **25**(3): p. 106-10.
5. Bates, D.O., et al., *Regulation of microvascular permeability by vascular endothelial growth factors.* J Anat, 2002. **200**(6): p. 581-97.
6. Burri, P.H., R. Hlushchuk, and V. Djonov, *Intussusceptive angiogenesis: its emergence, its characteristics, and its significance.* Dev Dyn, 2004. **231**(3): p. 474-88.
7. Ferrara, N., *Role of vascular endothelial growth factor in regulation of physiological angiogenesis.* Am J Physiol Cell Physiol, 2001. **280**(6): p. C1358-66.
8. Ferrara, N., *Vascular endothelial growth factor as a target for anticancer therapy.* Oncologist, 2004. **9 Suppl 1**: p. 2-10.
9. Ferrara, N., *VEGF: an update on biological and therapeutic aspects.* Curr Opin Biotechnol, 2000. **11**(6): p. 617-24.
10. Vincenti, V., et al., *Assignment of the vascular endothelial growth factor gene to human chromosome 6p21.3.* Circulation, 1996. **93**(8): p. 1493-5.
11. Houck, K.A., et al., *The vascular endothelial growth factor family: identification of a fourth molecular species and characterization of alternative splicing of RNA.* Mol Endocrinol, 1991. **5**(12): p. 1806-14.
12. Tischer, E., et al., *The human gene for vascular endothelial growth factor. Multiple protein forms are encoded through alternative exon splicing.* J Biol Chem, 1991. **266**(18): p. 11947-54.
13. Park, J.E., G.A. Keller, and N. Ferrara, *The vascular endothelial growth factor (VEGF) isoforms: differential deposition into the subepithelial extracellular matrix and bioactivity of extracellular matrix-bound VEGF.* Mol Biol Cell, 1993. **4**(12): p. 1317-26.
14. Ferrara, N., et al., *Heterozygous embryonic lethality induced by targeted inactivation of the VEGF gene.* Nature, 1996. **380**(6573): p. 439-42.
15. Carmeliet, P., et al., *Abnormal blood vessel development and lethality in embryos lacking a single VEGF allele.* Nature, 1996. **380**(6573): p. 435-9.
16. Ziche, M., S. Donnini, and L. Morbidelli, *Development of new drugs in angiogenesis.* Curr Drug Targets, 2004. **5**(5): p. 485-93.
17. Dor, Y., R. Porat, and E. Keshet, *Vascular endothelial growth factor and vascular adjustments to perturbations in oxygen homeostasis.* Am J Physiol Cell Physiol, 2001. **280**(6): p. C1367-74.

18. Garvin, S., et al., *Estradiol increases VEGF in human breast studied by whole-tissue culture.* Cell Tissue Res, 2006. **325**(2): p. 245-51.
19. Tran, D.C., et al., *Inhalant nitrite exposure alters mouse hepatic angiogenic gene expression.* Biochem Biophys Res Commun, 2003. **310**(2): p. 439-45.
20. Frank, S., et al., *Regulation of vascular endothelial growth factor expression in cultured keratinocytes. Implications for normal and impaired wound healing.* J Biol Chem, 1995. **270**(21): p. 12607-13.
21. Pertovaara, L., et al., *Vascular endothelial growth factor is induced in response to transforming growth factor-beta in fibroblastic and epithelial cells.* J Biol Chem, 1994. **269**(9): p. 6271-4.
22. Warren, R.S., et al., *Induction of vascular endothelial growth factor by insulin-like growth factor 1 in colorectal carcinoma.* J Biol Chem, 1996. **271**(46): p. 29483-8.
23. Gille, H., et al., *Analysis of biological effects and signaling properties of Flt-1 (VEGFR-1) and KDR (VEGFR-2). A reassessment using novel receptor-specific vascular endothelial growth factor mutants.* J Biol Chem, 2001. **276**(5): p. 3222-30.
24. Drake, C.J., et al., *VEGF regulates cell behavior during vasculogenesis.* Dev Biol, 2000. **224**(2): p. 178-88.
25. Joukov, V., et al., *Proteolytic processing regulates receptor specificity and activity of VEGF-C.* Embo J, 1997. **16**(13): p. 3898-911.
26. Kendall, R.L. and K.A. Thomas, *Inhibition of vascular endothelial cell growth factor activity by an endogenously encoded soluble receptor.* Proc Natl Acad Sci U S A, 1993. **90**(22): p. 10705-9.
27. Rahimi, N., *Vascular endothelial growth factor receptors: molecular mechanisms of activation and therapeutic potentials.* Exp Eye Res, 2006. **83**(5): p. 1005-16.
28. Ferrara, N., *Role of vascular endothelial growth factor in the regulation of angiogenesis.* Kidney Int, 1999. **56**(3): p. 794-814.
29. Hiratsuka, S., et al., *Involvement of Flt-1 tyrosine kinase (vascular endothelial growth factor receptor-1) in pathological angiogenesis.* Cancer Res, 2001. **61**(3): p. 1207-13.
30. Davis-Smyth, T., et al., *The second immunoglobulin-like domain of the VEGF tyrosine kinase receptor Flt-1 determines ligand binding and may initiate a signal transduction cascade.* Embo J, 1996. **15**(18): p. 4919-27.
31. Ferrara, N. and T. Davis-Smyth, *The biology of vascular endothelial growth factor.* Endocr Rev, 1997. **18**(1): p. 4-25.
32. Goldman, C.K., et al., *Paracrine expression of a native soluble vascular endothelial growth factor receptor inhibits tumor growth, metastasis, and mortality rate.* Proc Natl Acad Sci U S A, 1998. **95**(15): p. 8795-800.
33. Yano, S., et al., *Current status and perspective of angiogenesis and antivascular therapeutic strategy: non-small cell lung cancer.* Int J Clin Oncol, 2006. **11**(2): p. 73-81.
34. Kearney, J.B., et al., *Vascular endothelial growth factor receptor Flt-1 negatively regulates developmental blood vessel formation by modulating endothelial cell division.* Blood, 2002. **99**(7): p. 2397-407.
35. Fong, G.H., et al., *Role of the Flt-1 receptor tyrosine kinase in regulating the assembly of vascular endothelium.* Nature, 1995. **376**(6535): p. 66-70.

36. Barleon, B., et al., *Mapping of the sites for ligand binding and receptor dimerization at the extracellular domain of the vascular endothelial growth factor receptor FLT-1.* J Biol Chem, 1997. **272**(16): p. 10382-8.
37. Cunningham, S.A., et al., *Identification of the extracellular domains of Flt-1 that mediate ligand interactions.* Biochem Biophys Res Commun, 1997. **231**(3): p. 596-9.
38. Tanaka, K., et al., *Characterization of the extracellular domain in vascular endothelial growth factor receptor-1 (Flt-1 tyrosine kinase).* Jpn J Cancer Res, 1997. **88**(9): p. 867-76.
39. Kondo, K., et al., *Genomic organization of the flt-1 gene encoding for vascular endothelial growth factor (VEGF) receptor-1 suggests an intimate evolutionary relationship between the 7-Ig and the 5-Ig tyrosine kinase receptors.* Gene, 1998. **208**(2): p. 297-305.
40. Ito, T.K., et al., *Degradation of soluble VEGF receptor-1 by MMP-7 allows VEGF access to endothelial cells.* Blood, 2009. **113**(10): p. 2363-9.
41. Shibuya, M., *Vascular endothelial growth factor receptor-1 (VEGFR-1/Flt-1): a dual regulator for angiogenesis.* Angiogenesis, 2006. **9**(4): p. 225-30; discussion 231.
42. Baumann, M.U., et al., *First-trimester serum levels of soluble endoglin and soluble fms-like tyrosine kinase-1 as first-trimester markers for late-onset preeclampsia.* Am J Obstet Gynecol, 2008. **199**(3): p. 266 e1-6.
43. Artini, P.G., et al., *Vascular endothelial growth factor and its soluble receptor in patients with polycystic ovary syndrome undergoing IVF.* Hum Fertil (Camb), 2009. **12**(1): p. 40-4.
44. Tissot van Patot, M.C., et al., *Greater free plasma VEGF and lower soluble VEGF receptor-1 in acute mountain sickness.* J Appl Physiol, 2005. **98**(5): p. 1626-9.
45. Krussel, J.S., et al., *Expression of mRNA for vascular endothelial growth factor transmembraneous receptors Flt1 and KDR, and the soluble receptor sflt in cycling human endometrium.* Mol Hum Reprod, 1999. **5**(5): p. 452-8.
46. He, Y., et al., *Alternative splicing of vascular endothelial growth factor (VEGF)-R1 (FLT-1) pre-mRNA is important for the regulation of VEGF activity.* Mol Endocrinol, 1999. **13**(4): p. 537-45.
47. Clark, D.E., et al., *A vascular endothelial growth factor antagonist is produced by the human placenta and released into the maternal circulation.* Biol Reprod, 1998. **59**(6): p. 1540-8.
48. Martin, A., M.R. Komada, and D.C. Sane, *Abnormal angiogenesis in diabetes mellitus.* Med Res Rev, 2003. **23**(2): p. 117-45.
49. Folkman, J., *Angiogenesis in cancer, vascular, rheumatoid and other disease.* Nat Med, 1995. **1**(1): p. 27-31.
50. List, A.F., *Vascular endothelial growth factor signaling pathway as an emerging target in hematologic malignancies.* Oncologist, 2001. **6 Suppl 5**: p. 24-31.
51. Kerbel, R.S., *Tumor angiogenesis: past, present and the near future.* Carcinogenesis, 2000. **21**(3): p. 505-15.
52. Jain, R.K., *Tumor angiogenesis and accessibility: role of vascular endothelial growth factor.* Semin Oncol, 2002. **29**(6 Suppl 16): p. 3-9.

53. Arii, S., *Tumor angiogenesis and antiangiogenic therapy: current status and perspective.* Int J Clin Oncol, 2006. **11**(2): p. 71-2.
54. Ferrara, N., K.J. Hillan, and W. Novotny, *Bevacizumab (Avastin), a humanized anti-VEGF monoclonal antibody for cancer therapy.* Biochem Biophys Res Commun, 2005. **333**(2): p. 328-35.
55. Institute, N.C. *Clinical Trials: Vatalanib*. Available from: http://www.cancer.gov/search/ResultsClinicalTrials.aspx?protocolsearchid=6599893.
56. Ruggeri, B., et al., *CEP-7055: a novel, orally active pan inhibitor of vascular endothelial growth factor receptor tyrosine kinases with potent antiangiogenic activity and antitumor efficacy in preclinical models.* Cancer Res, 2003. **63**(18): p. 5978-91.
57. Jones-Bolin, S., et al., *The effects of the oral, pan-VEGF-R kinase inhibitor CEP-7055 and chemotherapy in orthotopic models of glioblastoma and colon carcinoma in mice.* Mol Cancer Ther, 2006. **5**(7): p. 1744-53.
58. Gomez-Manzano, C., et al., *VEGF Trap induces antiglioma effect at different stages of disease.* Neuro Oncol, 2008. **10**(6): p. 940-5.
59. Dean, M., *Approaches to identify genes for complex human diseases: lessons from Mendelian disorders.* Hum Mutat, 2003. **22**(4): p. 261-74.
60. Volm, M., et al., *Angiogenic growth factors and their receptors in non-small cell lung carcinomas and their relationships to drug response in vitro.* Anticancer Res, 1997. **17**(1A): p. 99-103.
61. Brown, L.F., et al., *Expression of vascular permeability factor (vascular endothelial growth factor) and its receptors in breast cancer.* Hum Pathol, 1995. **26**(1): p. 86-91.
62. Brown, L.F., et al., *Expression of vascular permeability factor (vascular endothelial growth factor) and its receptors in adenocarcinomas of the gastrointestinal tract.* Cancer Res, 1993. **53**(19): p. 4727-35.
63. Brown, L.F., et al., *Increased expression of vascular permeability factor (vascular endothelial growth factor) and its receptors in kidney and bladder carcinomas.* Am J Pathol, 1993. **143**(5): p. 1255-62.
64. Olson, T.A., et al., *Vascular permeability factor gene expression in normal and neoplastic human ovaries.* Cancer Res, 1994. **54**(1): p. 276-80.
65. Plate, K.H., et al., *Vascular endothelial growth factor is a potential tumour angiogenesis factor in human gliomas in vivo.* Nature, 1992. **359**(6398): p. 845-8.
66. Ferrara, N., et al., *Molecular and biological properties of the vascular endothelial growth factor family of proteins.* Endocr Rev, 1992. **13**(1): p. 18-32.
67. Strizzi, L., et al., *Vascular endothelial growth factor is an autocrine growth factor in human malignant mesothelioma.* J Pathol, 2001. **193**(4): p. 468-75.
68. Fan, F., et al., *Expression and function of vascular endothelial growth factor receptor-1 on human colorectal cancer cells.* Oncogene, 2005. **24**(16): p. 2647-53.
69. Siemeister, G., et al., *Two independent mechanisms essential for tumor angiogenesis: inhibition of human melanoma xenograft growth by interfering with either the vascular endothelial growth factor receptor pathway or the Tie-2 pathway.* Cancer Res, 1999. **59**(13): p. 3185-91.

70. Herbert, A. and A. Rich, *RNA processing and the evolution of eukaryotes.* Nat Genet, 1999. **21**(3): p. 265-9.
71. Twyman, R.M., *Advanced Molecular Biology : A Concise Reference* 1998: Oxford, UK BIOS Scientific Publishers, Ltd., 1998. .
72. Goldstrohm, A.C., A.L. Greenleaf, and M.A. Garcia-Blanco, *Co-transcriptional splicing of pre-messenger RNAs: considerations for the mechanism of alternative splicing.* Gene, 2001. **277**(1-2): p. 31-47.
73. Kornblihtt, A.R., et al., *Multiple links between transcription and splicing.* Rna, 2004. **10**(10): p. 1489-98.
74. Proudfoot, N., *Connecting transcription to messenger RNA processing.* Trends Biochem Sci, 2000. **25**(6): p. 290-3.
75. Krainer, A.R., *Eukaryotic mRNA Processing*. 1997: New York, Oxford University Press.
76. Nilsen, T.W., *The spliceosome: the most complex macromolecular machine in the cell?* Bioessays, 2003. **25**(12): p. 1147-9.
77. Cartegni, L., et al., *ESEfinder: a web resource to identify exonic splicing enhancers.* Nucleic Acid Research, 2003. **31**(13): p. 3568-3571.
78. Wahle, E. and U. Ruegsegger, *3'-End processing of pre-mRNA in eukaryotes.* FEMS Microbiol Rev, 1999. **23**(3): p. 277-95.
79. Grabowski, P.J. and D.L. Black, *Alternative RNA splicing in the nervous system.* Prog Neurobiol, 2001. **65**(3): p. 289-308.
80. Black, D.L., *Mechanisms of alternative pre-messenger RNA splicing.* Annu Rev Biochem, 2003. **72**: p. 291-336.
81. Horiuchi, T. and T. Aigaki, *Alternative trans-splicing: a novel mode of pre-mRNA processing.* Biol Cell, 2006. **98**(2): p. 135-40.
82. Shastry, B.S., *SNP alleles in human disease and evolution.* J Hum Genet, 2002. **47**(11): p. 561-6.
83. Botstein, D. and N. Risch, *Discovering genotypes underlying human phenotypes: past successes for mendelian disease, future approaches for complex disease.* Nat Genet, 2003. **33 Suppl**: p. 228-37.
84. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
85. Conde, L., et al., *PupaSNP Finder: a web tool for finding SNPs with putative effect at transcriptional level.* Nucleic Acids Res, 2004. **32**(Web Server issue): p. W242-8.
86. Halushka, M.K., et al., *Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.* Nat Genet, 1999. **22**(3): p. 239-47.
87. Shastry, B.S., *SNPs in disease gene mapping, medicinal drug development and evolution.* J Hum Genet, 2007.
88. Immervoll, T., et al., *Fine mapping and single nucleotide polymorphism association results of candidate genes for asthma and related phenotypes.* Hum Mutat, 2001. **18**(4): p. 327-36.
89. Kocsis, A.K., et al., *Association of beta-defensin 1 single nucleotide polymorphisms with Crohn's disease.* Scand J Gastroenterol, 2007: p. 1-9.
90. Wood, L.S., et al., *Association between neprilysin polymorphisms and sporadic Alzheimer's disease.* Neurosci Lett, 2007.

91. Zoledziewska, M., et al., *Further evidence of a primary, causal association of the PTPN22 620W variant with type 1 diabetes.* Diabetes, 2007.
92. Tsunoda, K., et al., *Single nucleotide polymorphism (D68D, T to C) in the syntaxin 1A gene correlates to age at onset and insulin requirement in Type II diabetic patients.* Diabetologia, 2001. **44**(11): p. 2092-7.
93. Donn, R.P., et al., *A novel 5'-flanking region polymorphism of macrophage migration inhibitory factor is associated with systemic-onset juvenile idiopathic arthritis.* Arthritis Rheum, 2001. **44**(8): p. 1782-5.
94. Kawasaki, A., et al., *Analysis on the association of human BLYS (BAFF, TNFSF13B) polymorphisms with systemic lupus erythematosus and rheumatoid arthritis.* Genes Immun, 2002. **3**(7): p. 424-9.
95. Ikeda, T., et al., *Association analysis of single nucleotide polymorphisms in cartilage-specific collagen genes with knee and hip osteoarthritis in the Japanese population.* J Bone Miner Res, 2002. **17**(7): p. 1290-6.
96. Lopes, C., et al., *PAI-1 polymorphisms modulate phenotypes associated with the metabolic syndrome in obese and diabetic Caucasian population.* Diabetologia, 2003. **46**(9): p. 1284-90.
97. Nyegaard, M., et al., *Novel polymorphisms in the somatostatin receptor 5 (SSTR5) gene associated with bipolar affective disorder.* Mol Psychiatry, 2002. **7**(7): p. 745-54.
98. Bewick, M.A., M.S. Conlon, and R.M. Lafrenie, *Polymorphisms in manganese superoxide dismutase, myeloperoxidase and glutathione-S-transferase and survival after treatment for metastatic breast cancer.* Breast Cancer Res Treat, 2007.
99. Gaasenbeek, M., et al., *Combined array-comparative genomic hybridization and single-nucleotide polymorphism-loss of heterozygosity analysis reveals complex changes and multiple forms of chromosomal instability in colorectal cancers.* Cancer Res, 2006. **66**(7): p. 3471-9.
100. Fredriksson, H., et al., *Identification of germline MLH1 alterations in familial prostate cancer.* Eur J Cancer, 2006. **42**(16): p. 2802-6.
101. Sun, T., et al., *Haplotypes in matrix metalloproteinase gene cluster on chromosome 11q22 contribute to the risk of lung cancer development and progression.* Clin Cancer Res, 2006. **12**(23): p. 7009-17.
102. Menendez, D., et al., *A SNP in the flt-1 promoter integrates the VEGF system into the p53 transcriptional network.* Proc Natl Acad Sci U S A, 2006. **103**(5): p. 1406-11.
103. ElSharawy, A., et al., *SNPSplicer: systematic analysis of SNP-dependent splicing in genotyped cDNAs.* Hum Mutat, 2006. **27**(11): p. 1129-34.
104. Vega, D.L., H. F.M. Isaac, and C.R. Scafe, *A Tool for Selecting SNPs for Association Studies Based on Observed Linkage Disequilibrium Patterns.* Pacific Symposium on Biocomputing, 2006. **11**: p. 487-498.
105. Cartegni, L., S.L. Chew, and A.R. Krainer, *Listening to silence and understanding nonsense: exonic mutations that affect splicing.* Nat Rev Genet, 2002. **3**(4): p. 285-98.
106. Ishii, T.M., et al., *Site-directed mutagenesis.* Methods Enzymol, 1998. **293**: p. 53-71.

107. Roche, R.I., *Role of RNA Processing Factors in the Expression of Flt-1 and its Secreted Variant, sFlt-1* in *Department of Biomedical Sciences and Pathobiology*. 2005, Virginia Polytechnic Institute and
State University: Blacksburg. p. 201.
108. Cooper, T.A., *Use of minigene systems to dissect alternative splicing elements.* Methods, 2005. **37**(4): p. 331-40.
109. Stephan, D.A. and S.B. Glueck, *Containing multitudes: focus on "novel and nondetected human signaling protein polymorphisms".* Physiol Genomics, 2002. **10**(3): p. 127-9.
110. Fairbrother, W.G. and L.A. Chasin, *Human genomic sequences that inhibit splicing.* Mol Cell Biol, 2000. **20**(18): p. 6816-25.
111. Lei, H. and I. Vorechovsky, *Identification of splicing silencers and enhancers in sense Alus: a role for pseudoacceptors in splice site repression.* Mol Cell Biol, 2005. **25**(16): p. 6912-20.
112. Sun, H. and L.A. Chasin, *Multiple splicing defects in an intronic false exon.* Mol Cell Biol, 2000. **20**(17): p. 6414-25.
113. Cartegni, L. and A.R. Krainer, *Correction of disease-associated exon skipping by synthetic exon-specific activators.* Nat Struct Biol, 2003. **10**(2): p. 120-5.
114. D'Souza, I., et al., *Missense and silent tau gene mutations cause frontotemporal dementia with parkinsonism-chromosome 17 type, by affecting multiple alternative RNA splicing regulatory elements.* Proc Natl Acad Sci U S A, 1999. **96**(10): p. 5598-603.
115. Shiga, N., et al., *Disruption of the splicing enhancer sequence within exon 27 of the dystrophin gene by a nonsense mutation induces partial skipping of the exon and is responsible for Becker muscular dystrophy.* J Clin Invest, 1997. **100**(9): p. 2204-10.
116. Lorson, C.L., et al., *A single nucleotide in the SMN gene regulates splicing and is responsible for spinal muscular atrophy.* Proc Natl Acad Sci U S A, 1999. **96**(11): p. 6307-11.
117. Wang, Z., et al., *General and specific functions of exonic splicing silencers in splicing control.* Mol Cell, 2006. **23**(1): p. 61-70.
118. Sikorski, R. and R. Peters, *SOFTWARE: Digital Duct Tape.* Science 1999. **284**(5418): p. 1.
119. Akerman, M., et al., *A computational approach for genome-wide mapping of splicing factor binding sites.* Genome Biol, 2009. **10**(3): p. R30.
120. Goodman, A.H., D. Heath, and M.S. Lindee, *Genetic Nature/culture : Anthropology and Science Beyond the Two-culture Divide* 2003, Berkeley University of California Press.
121. Page, R.D.M. and E.C. Holmes, *Molecular Evolution : A Phylogenetic Approach* 1998, Oxford ; Malden, Mass. Blackwell Science.
122. Albertson, T.M. and B.D. Preston, *DNA replication fidelity: proofreading in trans.* Curr Biol, 2006. **16**(6): p. R209-11.
123. Terwilliger, J.D., et al., *Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion.* Hum Hered, 1998. **48**(3): p. 138-54.

124. Nielsen, R., et al., *Recent and ongoing selection in the human genome.* Nat Rev Genet, 2007. **8**(11): p. 857-68.
125. Thomas, J.W., et al., *Comparative analyses of multi-species sequences from targeted genomic regions.* Nature, 2003. **424**(6950): p. 788-93.
126. Soller, J.T., et al., *Comparison of the human and canine cytokines IL-1(alpha/beta) and TNF-alpha to orthologous other mammalians.* J Hered, 2007. **98**(5): p. 485-90.
127. Petit, V., et al., *Downstream-of-FGFR is a fibroblast growth factor-specific scaffolding protein and recruits Corkscrew upon receptor activation.* Mol Cell Biol, 2004. **24**(9): p. 3769-81.
128. Goodman, S.J., et al., *Alternative splicing affecting a novel domain in the C. elegans EGL-15 FGF receptor confers functional specificity.* Development, 2003. **130**(16): p. 3757-66.
129. Mayor, C., et al., *VISTA : visualizing global DNA sequence alignments of arbitrary length.* Bioinformatics, 2000. **16**(11): p. 1046-7.
130. Brudno, M., et al., *LAGAN and Multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA.* Genome Res, 2003. **13**(4): p. 721-31.
131. Kappas, N.C., et al., *The VEGF receptor Flt-1 spatially modulates Flk-1 signaling and blood vessel branching.* J Cell Biol, 2008. **181**(5): p. 847-58.
132. Zeng, G., et al., *Orientation of endothelial cell division is regulated by VEGF signaling during blood vessel formation.* Blood, 2007. **109**(4): p. 1345-52.
133. Sugino, N., et al., *Expression of vascular endothelial growth factor (VEGF) receptors in rat corpus luteum: regulation by oestradiol during mid-pregnancy.* Reproduction, 2001. **122**(6): p. 875-81.
134. Mochida, S., et al., *Increased expressions of vascular endothelial growth factor and its receptors, flt-1 and KDR/flk-1, in regenerating rat liver.* Biochem Biophys Res Commun, 1996. **226**(1): p. 176-9.
135. Li, J., et al., *VEGF, flk-1, and flt-1 expression in a rat myocardial infarction model of angiogenesis.* Am J Physiol, 1996. **270**(5 Pt 2): p. H1803-11.
136. Cai, J., et al., *Activation of vascular endothelial growth factor receptor-1 sustains angiogenesis and Bcl-2 expression via the phosphatidylinositol 3-kinase pathway in endothelial cells.* Diabetes, 2003. **52**(12): p. 2959-68.
137. Cai, J., et al., *Pigment epithelium-derived factor inhibits angiogenesis via regulated intracellular proteolysis of vascular endothelial growth factor receptor 1.* J Biol Chem, 2006. **281**(6): p. 3604-13.
138. Uchida, N., et al., *Distribution of VEGF and flt-1 in the normal dog tissues.* J Vet Med Sci, 2008. **70**(11): p. 1273-6.
139. Illumina. Available from: http://www.illumina.com/pages.ilmn?ID=203.
140. AppliedBiosystems. Available from: http://www3.appliedbiosystems.com/AB_Home/applicationstechnologies/SOLiDSystemSequencing/index.htm.

Appendix A: Materials and Methods (Chapters 2 and 3)

Polymerase Chain Reaction (PCR)
All PCR reactions were performed on the Hybaid Limited Thermalcycler #SPRT001, Issue 2

Standard reaction conditions: with 0.25 µM for each primer and 10 µL of 1x Taq PCR Master Mix (QIAGEN, 201443), supplemented with nuclease free water for a 20 µL reaction.

PCR samples were either run on an agarose gel of appropriate concentration according to the fragment size, and then gel purified from the gel using the QIAEX® II Gel Extraction Kit (Catalog #20021) or desalted from solution with the same kit, using the modified protocol indicated in the QIAEX II booklet.

Standard PCR was run as follows:

Stage	Step	Temperature (°C)	Time (minutes)	# Cycles
1	1	94	2	1
	2	60	1	
	3	72	2	
	4	0	0	
2	1	94	1	28
	2	60	1	
	3	72	2	
	4	0	0	
3	1	94	1	1
	2	60	1	
	3	72	10	
	4	0	0	
4	1	4	Hold	N/A
			Total Cycles:	30

Mutant PCR (1° and 2° PCR) was run as follows:

Stage	Step	Temperature (°C)	Time (minutes)	# Cycles
1	1	94	2	1
	2	55	1	
	3	72	2	
	4	0	0	
2	1	94	1	18
	2	55	1	
	3	72	2	
	4	0	0	
3	1	94	1	1
	2	55	1	
	3	72	10	
	4	0	0	
4	1	4	Hold	N/A
			Total Cycles:	20

Agarose Gel Electrophoresis and Gel Purification

Samples to be run on gel were supplemented with Gel Loading Solution (0.3% of the sample) and run in a gel electrophoresis apparatus at 100V for 30-45 minutes (depending on size of the fragments indicated) in 1X TBE running buffer. Gels were stained with Ethidium Bromide (1.5 µg/ml in ddH2O) for 20 minutes and de-stained in ddH2O for an additional 20 minutes. Bands on the gel were visualized via an ultra-violet transilluminating light box, using molecular markers (λ-*Hind*III plus φX174-*Hae*III) for band size separation.

Running Buffer (1x TBE from 10x)
0.89M Tris
0.87M Boric Acid
0.027M EDTA

Gel Loading Solution (Sigma Gel Loading Solution, Catalog #G2526)
0.05% (w/v) Bromophenol blue
40% (w/v) Sucrose
0.1 M EDTA, pH 8.0
0.5% (w/v) Sodium lauryl sulfate (SDS)

Marker Mix for agarose gel electrophroesis: (New England Biolabs*)*
λ DNA- *Hind*III Digest (#N3012) 0.5µg/25µl final concentration
ΦX174 DNA – *Hae*III Digest (#N3026) 0.5µg/25 µL final concentration

The two components were mixed with Sigma Gel Loading Solution (#G2526) (1 volume dye/3 volumes sample) and nuclease-free water to 25 µL final volume.

Cloning

Preparation of clones involved several steps. Once the insert was prepared (by amplifying relevant insert from template or restriction enzyme digest to excise, as required for experiment) and the plasmid linearized by suitable restriction enzyme digest, both were run on the appropriate concentration of agarose gel and purified from the gel using QIAEX® II Gel Extraction Kit (Catalog #20021). In some cases of linearization, the plasmid could be directly desalted from solution (without running on a gel) by utilizing the QIAEX® II modified protocol included with the kit.

Once the insert and plasmid were purified, they were ligated in an 8:2 or 9:3 molar ratio (for larger plasmids) with T4 DNA Ligase (1 unit/µl, Catalog #15224-017) supplemented with 4X T4 DNA Ligase Buffer (supplied with T4 DNA Ligase) and left at room temperature overnight. Ligated plasmids were then transfected into chemically competent cells according to the manufacturers recommendation (ice incubation time, heat shock time, amount of S.O.C. medium to add, etc) and placed in a 37°C shaking incubator for one hour. One hundred µL of this mixture was then spread onto a 100 mm LB-amp plate (Luria-Bertani [LB] agar (Becton Dickinson # 244520) plus ampicillin at 0.1 mg/ml), and plates incubated overnight in a 37°C incubator. The next morning, colonies were picked from each plate and grown in a 1 mL mixture of LB broth (Becton Dickinson # 244620) plus ampicillin at 0.1 mg/ml, which were then left overnight in a in a 37°C shaking incubator. After 12-14 hours, 200 µL of each 1mL culture was transferred to a 1.5 ml tube and centrifuged for 5 minutes at room temperature (14,000 x g) and supernatant was aspirated. Cells were then resuspended in 100 µL 1x TE and heated at 95°C for 5 minutes. One µL of the TE lysate was then used to screen clones via standard PCR, and PCR products were then run on an agarose whereby orientation of clones as well as content could be determined. In order to confirm characterization of clone, appropriate bands were gel purified and sequenced in an automated sequencer (see below for details of sequencing).

Once a clone had been characterized, the plasmid could then be purified. This was done by streaking the positive clone onto an LB-amp clone and placed in a 37C incubator overnight. The next day, a single colony was picked off streaked plate and grown in a 100mL LB-amp culture overnight in a 37°C shaking (250rpm) incubator overnight. A 400 µL sample from the 100 ml culture was mixed with 400 µL of 50% glycerol and frozen at –70^{o}C for archival purposes. The rest was used to purify the plasmid.

All clones in the holding vector pL38 were purified with QIAGEN® Plasmid Purification Midi Kit (Catalog #12143). All clones in the expression vector pcDNAIntA were purified with the QIAGEN® EndoFree Plasmid Purification Maxi Kit (Catalog #12362). However, once all mutant clones had been assembled, all mutants and the wild type clone were re-grown from bacterial aliquots, streaked onto fresh plates and purified with the Zyppy Midiprep kit (Zymo Research, Catalog #D4025), and it is these Zyppy-purified clones were the ones used for transfection experiments

After purification of plasmid, the pellet was redissolved in 200 µL of nuclease-free H2O when using the QIAGEN preps, or 150µl of nuclease-free H2O when using the Zyppy preps. Concentration and purity for QIAGEN plasmids was estimated by absorbance readings at 260 nm and 280 nm ([DNA] µg/ml = A260 x dilution factor x 50 µg/ml [extinction coefficient]). Concentration and purity for Zyppy preps was quantified with the NanoDrop 1000 Spectrophotometer (Thermo Scientific Inc., Wilmington, DE, USA).

Competent cells used in this study included the following:
1. FusionBlue Competent Cells (Clontech, Catalog #636700)
2. DH5α cells (Subcloning Efficiency, Invitrogen, Catalog #18265-017)
3. OneShot TOP10 cells (Invitrogen, Catalog #C3030-03)

Cell Culture

NIH/3T3 mouse fibroblast cells (ATCC #CRL-1658) were grown as per ATCC Guidelines in DMEM/FBS/Gent, in a humidified incubator at 37°C and 5% CO_2. Medium was replaced every two days, and each flask was subcultured when cells reached 70-80% confluence. Separate flasks were grown in order to ensure that each transfection experiment arose out of a distinct population of cells and, in most cases, a different passage number.

Passaging cells involved removing and discarding the culture medium, rinsing the flask with 10 mL DPBS to remove traces of medium and adding 2 mLs of 0.25% (w/v) Trypsin to flask. Once cells were detached (10-15 minutes), 8 mLs of complete growth medium was added and after mixing thoroughly, the cell suspension was transferred to a 50 ml conical bottom tube. A 100 µL aliquot of cells was mixed with 100 µL of 0.4% Trypan blue in DPBS for counting cells under the microscope using a hemacytometer. The remaining cell suspension was centrifuged at 150 x g in a clinical centrifuge for 10 minutes. After aspirating supernatant, cells were resuspended in a volume of fresh DMEM/FBS/Gent appropriate to inoculate flasks or dishes at 5 X 10^3 cells/cm^2. Cells were supplemented with growth medium up to 10 mLs to be placed back into a T75 flask. If setting up a transfection experiments, cells were inoculated at 6 X 10^3 cells/cm^2 into eight 60mm dishes (Corning) and incubated for at least 24 hours before proceeding with transfection.

DMEM/FBS/Gent:
Dulbecco's Modified Eagle Medium (DMEM) with 4.5 mg/ml glucose,
glutamine (Mediatech #10-013-CV), supplemented with 10% (v/v) Fetal Bovine Serum (FBS) (Mediatech #14-501-F) and 50 µg/ml Gentamycin (LTI #15710-072).

DPBS: Dulbecco's Phosphate Buffered Saline, Ca2+/Mg2+-free (Mediatech #21-031-CM).

Trypsin (1X): (Mediatech #25-053-CI)

Transfection, RNA Isolation/Purification and cDNA Synthesis
Zyppy purified plasmids were transfected into 3T3/NIH cells using a liposomal formulation (Mirus TransIT®-3T3, Catalog# MIR 2180) in 60mm Corning dishes at 50-60% cell confluence. This was done by preparing the optimized amounts of the Mirius transfection reagents (15uL of TransIT, 5uL of the 3T3 Authority Agent) into 250 µL of OPTI-MEM-I media (Invitrogen, Catalog # 31985-070) in a sterile tube, and mixed with 3 µg of plasmid DNA and mixtures were left at room temperature for 20-30 minutes. The medium was replaced on the cells, and then the transfection mixtures added drop-wise to the corresponding dish. Dishes were rocked back and forth to ensure even distribution of transfection mixes, and returned to the incubator.

After 48 hours had passed, transfected cells were ready to be harvested, which was done by removal of media from dish and addition of 600uL of Buffer RLT (containing guanidine thiocyanate [from QIAGEN RNeasy® Mini Kit, Catalog #74104]) plus 1% β-mercaptoethanol. Cells were scraped off the surface of the dish using a Costar plastic cell lifter (Catalog #EF8612A), and lysate was t transferred to a centrifugal homogenizer (QIAshredder, QIAGEN, Catalog #79654). QIAshredders were centrifuged at 14,000 x *g* for 2 minutes, and the process of RNA isolation was continued with the QIAGEN RNeasy® Mini Kit (Catalog #74104). RNA was eluted with 2 applications of 40 µL of nuclease-free water. Once the RNA had been isolated, DNaseI digestion was performed (RNase-Free DNase Set, QIAGEN, Catalog #79254) to further reduce carryover of plasmid DNA. DNaseI-digested RNA lysates were then purified using the protocol indictaed in the RNeasy protocol kit, and eluted from membrane with 40 µL of nuclease free water. RNA was quantified using the NanoDrop.

cDNA was synthesized using High Capacity cDNA Reverse Transcription Kit (Applied Biosystems, Catalog #4368814) as indicated in the manufacturers protocol, by combining 10 µL of master mix with 10 µL of RNA (concentrations ranged from 30.5-296 ng/µL). Samples were then placed in the thermocycler for 10 minutes at 25°C, 120 minutes @ 37°C and 5 seconds @85°C, with a holding step of 4°C.

Separate samples were prepared containing 10 µL total RNA combined with 10 µL nuclease free water for non-reverse transcribed (No RT) controls.

Real-Time Quantitative Polymerase Chain Reaction (QPCR)
Previous work in the laboratory [127] established QPCR reagents, protocols and primers that reliably detect Flt1 or sFlt1 mRNA, as well as the neomycin resistance gene (Neo^R) which was used for normalizing the results. This previous work also determined the selectivity of the primers used to detect human Flt1 and sFlt1 and not the mouse variants of Flt1/sFlt due to the lack of a complete match of mouse Flt1/sFlt1 with the primers [127].

The QPCR experiments in this study utilized the SYBR®Green PCR Master Mix (PE Biosystems,Catalog #4309155), and each reaction mixture consisted of 42 µL of SYBR Green Master Mix, 42 µL of 5' and 3' primer at 10 uM each (0.30uM total) and 1uL of template cDNA. Once mixes had been prepared, each reaction mix was distributed into 3 wells of a 96 well

optical PCR plate for a triplicate of each sample. For each of the sFlt and Neo samples, a "No RT" control was set up and run in triplicate alongside the cDNA sample, which allowed for correction of background noise from any carryover plasmid that could contribute to the signals detected. For Flt1, only a subset of the "No RT" samples were run as a control, as Flt1 cDNA does not exist in the NoRT product.

The Neo^R gene, which lies on the pcDNAIntA vector and is expressed under a separate promoter, was utilized for normalization as it can help eliminate variation observed from conditions not related to the reaction, such as user handling and differential transfection efficiencies among the test plasmids used.

Samples were then put into the ABI Prism® Sequence Detection System (SDS) 7300. Reactions were run under the following conditions: 1 cycle at stage 1 (50^{o}C for 2 minutes), 1 cycle at stage 2 (95^{o}C, 10 minutes), 40 cycles at (stage 3, 95^{o}C, 15 seconds) and 1 cycle at stage 4(60^{o}C, 1 minute). This was followed by a dissociation stage (95^{o}C for 15 second, 60^{o}C for 1 minute and 95^{o}C for 15 seconds). The dissociation curves were checked for one measured peak denoting primer/template interaction.

The ABI software determines a threshold cycle (Ct) at which a product is measurable over noise. The smaller the Ct number, the more template was present at the start of the reaction, as the amplified product was detected with a fewer number of cycles. Flt1 and sFlt1 mRNA expression levels (normalized internally to Neo^R mRNA) relative to levels in wild type transfectants were calculated by the "ΔΔCt" method. The ΔCt is the difference between Flt1 (or sFlt1) Ct and Neo^R Ct for a given sample (corrected for plasmid background). The ΔΔCt is the difference between ΔCts for a given mutant sample and that occurring in the wild type. Ratios to wild type were then calculated as $2^{-\Delta\Delta Ct}$ to account for a two-fold difference in target concentration per cycle.

In addition, Flt1 and sFlt1 mRNA were also compared to each other, measuring relative levels. This was done by calculating the ΔCt (in this case, the difference between Flt1 and sFlt1 Ct for a given sample), followed by the ΔΔCt (the difference between ΔCts for a given mutant sample and that occurring in the wild type). Ratios to wild type were then calculated as $2^{-\Delta\Delta Ct}$ to account for a two-fold difference in target concentration per cycle.

DNA Sequencing

For all reactions in Chapter 2, automated DNA sequencing was performed in the Core Laboratory Facility (CLF) at the Virginia Bioinformatics Institute (VBI) using standard methods on an ABI 377 automated DNA Sequencer or an ABI 3100 capillary sequencer.

For the human genomic samples in Chapter 3, preliminary testing of primers and samples was performed in the CLF at VBI as indicated above, but the entire genomic SNP screen was sequenced at Wake Forest University Biomolecular Resource Laboratory CLF using standard methods on an Applied Biosystems Model 3100 Genetic Analyzer.

Appendix B: Statistical Data (Chapter 2)

FLt_Neo 1-4: SNP, DBL, and WT: Anova to compare SNP, DBL, and WT clones

The Mixed Procedure

Model Information	
Data Set	WORK.ONEFOUR2
Dependent Variable	Flt_Neo
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

Class Level Information		
Class	**Levels**	**Values**
Clone	5	Dbl ESE Dbl SNP E13 SNP E14 SNP WT

Dimensions	
Covariance Parameters	1
Columns in X	6
Columns in Z	0
Subjects	1
Max Obs Per Subject	19

Number of Observations	
Number of Observations Read	19
Number of Observations Used	19
Number of Observations Not Used	0

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Residual	0.7908

FLt_Neo 1-4: SNP, DBL, and WT: Anova to compare SNP, DBL, and WT clones

The Mixed Procedure

Fit Statistics	
-2 Res Log Likelihood	43.1
AIC (smaller is better)	45.1
AICC (smaller is better)	45.4
BIC (smaller is better)	45.7

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Clone	4	14	6.08	0.0047

Least Squares Means									
Effect	Clone	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Clone	Dbl ESE	-6.8900	0.5134	14	-13.42	<.0001	0.05	-7.9911	-5.7889
Clone	Dbl SNP	-5.2584	0.4446	14	-11.83	<.0001	0.05	-6.2120	-4.3048
Clone	E13 SNP	-4.4895	0.4446	14	-10.10	<.0001	0.05	-5.4431	-3.5359
Clone	E14 SNP	-4.6000	0.4446	14	-10.35	<.0001	0.05	-5.5536	-3.6464
Clone	WT	-6.7121	0.4446	14	-15.10	<.0001	0.05	-7.6657	-5.7585

Differences of Least Squares Means												
Effect	Clone	Clone	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper
Clone	Dbl ESE	WT	-0.1779	0.6792	14	-0.26	0.7972	Dunnett	0.9966	0.05	-1.6346	1.2788
Clone	Dbl SNP	WT	1.4537	0.6288	14	2.31	0.0365	Dunnett	0.1121	0.05	0.1050	2.8023
Clone	E13 SNP	WT	2.2226	0.6288	14	3.53	0.0033	Dunnett	0.0113	0.05	0.8740	3.5712
Clone	E14 SNP	WT	2.1121	0.6288	14	3.36	0.0047	Dunnett	0.0158	0.05	0.7635	3.4607

Differences of Least Squares Means				
Effect	Clone	Clone	Adj Lower	Adj Upper
Clone	Dbl ESE	WT	-2.0495	1.6937
Clone	Dbl SNP	WT	-0.2791	3.1865
Clone	E13 SNP	WT	0.4898	3.9554
Clone	E14 SNP	WT	0.3793	3.8449

sFLt_Neo 1-4: SNP, DBL, and WT: Anova to compare SNP, DBL, and WT clones

The Mixed Procedure

Model Information	
Data Set	WORK.ONEFOUR2
Dependent Variable	sFlt_Neo
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

Class Level Information		
Class	**Levels**	**Values**
Clone	5	Dbl ESE Dbl SNP E13 SNP E14 SNP WT

Dimensions	
Covariance Parameters	1
Columns in X	6
Columns in Z	0
Subjects	1
Max Obs Per Subject	19

Number of Observations	
Number of Observations Read	19
Number of Observations Used	19
Number of Observations Not Used	0

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Residual	1.1899

sFLt_Neo 1-4: SNP, DBL, and WT: Anova to compare SNP, DBL, and WT clones

The Mixed Procedure

Fit Statistics	
-2 Res Log Likelihood	48.8
AIC (smaller is better)	50.8
AICC (smaller is better)	51.1
BIC (smaller is better)	51.4

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Clone	4	14	0.59	0.6756

Least Squares Means									
Effect	Clone	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Clone	Dbl ESE	-7.5447	0.6298	14	-11.98	<.0001	0.05	-8.8954	-6.1939
Clone	Dbl SNP	-8.4691	0.5454	14	-15.53	<.0001	0.05	-9.6389	-7.2993
Clone	E13 SNP	-7.7755	0.5454	14	-14.26	<.0001	0.05	-8.9453	-6.6057
Clone	E14 SNP	-8.0689	0.5454	14	-14.79	<.0001	0.05	-9.2387	-6.8991
Clone	WT	-8.5786	0.5454	14	-15.73	<.0001	0.05	-9.7484	-7.4088

Differences of Least Squares Means												
Effect	Clone	Clone	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper
Clone	Dbl ESE	WT	1.0339	0.8331	14	1.24	0.2350	Dunnett	0.5635	0.05	-0.7530	2.8208
Clone	Dbl SNP	WT	0.1095	0.7713	14	0.14	0.8891	Dunnett	0.9997	0.05	-1.5448	1.7638
Clone	E13 SNP	WT	0.8031	0.7713	14	1.04	0.3154	Dunnett	0.6947	0.05	-0.8513	2.4574
Clone	E14 SNP	WT	0.5097	0.7713	14	0.66	0.5195	Dunnett	0.9097	0.05	-1.1447	2.1640

Differences of Least Squares Means				
Effect	Clone	Clone	Adj Lower	Adj Upper
Clone	Dbl ESE	WT	-1.2620	3.3298
Clone	Dbl SNP	WT	-2.0161	2.2351
Clone	E13 SNP	WT	-1.3225	2.9287
Clone	E14 SNP	WT	-1.6159	2.6353

FLt_sFLt 1-4: SNP, DBL, and WT: Anova to compare SNP, DBL, and WT clones

The Mixed Procedure

Model Information	
Data Set	WORK.ONEFOUR2
Dependent Variable	Flt_sFlt
Covariance Structure	Diagonal
Estimation Method	REML
Residual Variance Method	Profile
Fixed Effects SE Method	Model-Based
Degrees of Freedom Method	Residual

Class Level Information		
Class	**Levels**	**Values**
Clone	5	Dbl ESE Dbl SNP E13 SNP E14 SNP WT

Dimensions	
Covariance Parameters	1
Columns in X	6
Columns in Z	0
Subjects	1
Max Obs Per Subject	19

Number of Observations	
Number of Observations Read	19
Number of Observations Used	19
Number of Observations Not Used	0

Covariance Parameter Estimates	
Cov Parm	**Estimate**
Residual	0.8196

FLt_sFLt 1-4: SNP, DBL, and WT: Anova to compare SNP, DBL, and WT clones

The Mixed Procedure

Fit Statistics	
-2 Res Log Likelihood	43.6
AIC (smaller is better)	45.6
AICC (smaller is better)	45.9
BIC (smaller is better)	46.2

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Clone	4	14	6.07	0.0048

Least Squares Means									
Effect	Clone	Estimate	Standard Error	DF	t Value	Pr > |t|	Alpha	Lower	Upper
Clone	Dbl ESE	0.6547	0.5227	14	1.25	0.2309	0.05	-0.4664	1.7757
Clone	Dbl SNP	3.2107	0.4527	14	7.09	<.0001	0.05	2.2398	4.1815
Clone	E13 SNP	3.2860	0.4527	14	7.26	<.0001	0.05	2.3151	4.2569
Clone	E14 SNP	3.4689	0.4527	14	7.66	<.0001	0.05	2.4981	4.4398
Clone	WT	1.8665	0.4527	14	4.12	0.0010	0.05	0.8956	2.8374

Differences of Least Squares Means												
Effect	Clone	Clone	Estimate	Standard Error	DF	t Value	Pr > |t|	Adjustment	Adj P	Alpha	Lower	Upper
Clone	Dbl ESE	WT	-1.2118	0.6914	14	-1.75	0.1015	Dunnett	0.2826	0.05	-2.6948	0.2712
Clone	Dbl SNP	WT	1.3442	0.6402	14	2.10	0.0544	Dunnett	0.1617	0.05	-0.02884	2.7172
Clone	E13 SNP	WT	1.4195	0.6402	14	2.22	0.0436	Dunnett	0.1322	0.05	0.04650	2.7925
Clone	E14 SNP	WT	1.6024	0.6402	14	2.50	0.0253	Dunnett	0.0796	0.05	0.2294	2.9754

Differences of Least Squares Means				
Effect	Clone	Clone	Adj Lower	Adj Upper
Clone	Dbl ESE	WT	-3.1173	0.6936
Clone	Dbl SNP	WT	-0.4199	3.1083
Clone	E13 SNP	WT	-0.3446	3.1836
Clone	E14 SNP	WT	-0.1617	3.3665

Appendix C: IRB Forms (Chapter 3)

1. IRB Proposal

VirginiaTech

Institutional Review Board
Research Protocol

Email completed form and all applicable supporting documents (see, http://www.irb.vt.edu/pages/researchers.htm#supporting) to irb@vt.edu (PDFs preferred). Copy all investigators listed on the application.

Section 1: Project Information

1.1 PROJECT TITLE:

Screening Flt1 Region of Interest for SNPs

1.2 PRINCIPAL INVESTIGATOR (Virginia Tech faculty only)

Name: **Dr. William Huckle**	Virginia Tech department: **DBSP/ Vet Med**
Email address: **wrhuckle@vt.edu**	VT PID: wrhuckle
Human subject protections training (required, see http://www.irb.vt.edu/pages/training.htm) completed through: ☒ VT IRB Training ☐ Other, certificate is attached ☐ Other, training is on file with IRB office	

1.3 CO-INVESTIGATORS (Include any individual responsible for the design and conduct of the study, or who will use data for publication purposes. Attach separate pages as necessary.)

Name: **Burouj Ajlouni**	
Email address: **bajlouni@vt.edu**	VT PID -or- organization name if non-VT employee or agent: **bajlouni**
Human subject protections training (required) completed through: ☒ VT IRB Training ☐ Other, certificate is attached ☐ Other, training is on file with IRB office	

Name:	
Email address:	VT PID -or- organization name if non-VT employee or agent:
Human subject protections training (required) completed through: ☐ VT IRB Training ☐ Other, certificate is attached ☐ Other, training is on file with IRB office	

Name:	
Email address:	VT PID -or- organization name if non-VT employee or agent:
Human subject protections training (required) completed through: ☐ VT IRB Training ☐ Other, certificate is attached ☐ Other, training is on file with IRB office	

Include with this application a current CV, resume, or biosketch for all investigators listed above

1.4 DEPARTMENTAL REVIEWER

If required, view http://www.irb.vt.edu/pages/researchers.htm - click on "instructions" under Research Protocol.

Name:	VT PID:	Email:
_____ Signature	_____ Date	Mail (campus code: 0497), fax (540-231-0959), or scan/email (irb@vt.edu) signed page to the IRB office.

1.5 DO ANY OF THE INVESTIGATORS OF THIS PROJECT HAVE A REPORTABLE CONFLICT OF INTEREST? (http://www.irb.vt.edu/pages/researchers.htm#conflict)

☒ **No**
☐ **Yes,** explain:

1.6 WILL THIS RESEARCH INVOLVE COLLABORATION WITH ANOTHER INSTITUTION?

☒ **No,** go to question 1.7
☐ **Yes,** answer questions within table →

IF YES
Provide the name of the institution [for institutions located overseas, please also provide name of country]**:**
Indicate the status of this research project with the other institution's IRB: ☐ Pending approval ☐ Approved ***[include approval letter with protocol]*** ☐ Other institution does not have a human subject protections review board ☐ Other, explain:
Will the collaborating institution(s) be engaged in the research? (http://www.hhs.gov/ohrp/humansubjects/assurance/engage.htm) ☐ No ☐ Yes
Will Virginia Tech's IRB review all human subject research activities involved with this project? ☐ No, provide the name of the primary institution: ☐ Yes *Note: primary institution = primary recipient of the grant or main coordinating center*

1.7 IS THIS RESEARCH FUNDED?

☐ **No,** go to question 1.8
☒ **Yes,** answer questions within table →

IF YES
Provide the name of the sponsor [if NIH, specify department]: (VT various)
Provide the VT Office of Sponsored Programs (OSP) proposal, grant, or fund number related to this project: 445076 ☐ OSP number pending ☐ Submission through OSP not required for this project
Is this project receiving federal funds? ☒ No ☐ Yes ***[include grant application, OSP proposal, or "statement of work" with protocol]*** **If yes,** **Does the grant application, OSP proposal, or "statement of work" related to this project include activities involving human subjects that are <u>not</u> covered within this IRB application?** ☐ No, all human subject activities are covered in this IRB application

☐ Yes, however these activities will be covered in future VT IRB applications, these activities include:
☐ Yes, however these activities have been covered in past VT IRB applications, the IRB number(s) are as follows:
☐ Yes, however these activities have been or will be reviewed by another institution's IRB, the name of this institution is as follows:
☐ Other, explain:

Is Virginia Tech the primary awardee or the coordinating center of this grant?
☐ No, provide the name of the primary institution:
☐ Yes

1.8 DOES THIS STUDY *ONLY* INVOLVE THE COLLECTION OR STUDY OF EXISTING DATA?
Please note: it is not considered existing data if a researcher transfers to Virginia Tech from another institution and will be conducting data analysis of an on-going study.

☒ **No,** go to question 1.9
☐ **Yes,** respond only to the following sections within this document: **Section 1** (Project Information), **Section 2** (Justification), **Section 8** (Confidentiality/Anonymity), and **Section 14** (Research Involving Existing Data)

1.9 DOES THIS STUDY INVOLVE CONFIDENTIAL OR PROPRIETARY INFORMATION (OTHER THAN HUMAN SUBJECT CONFIDENTIAL INFORMATION), OR INFORMATION RESTRICTED FOR NATIONAL SECURITY OR OTHER REASONS BY A U.S. GOVERNMENT AGENCY?
For example – government / industry proprietary or confidential trade secret information

☒ **No**
☐ **Yes,** describe:

1.10 DOES THIS STUDY INVOLVE SHIPPING ANY TANGIBLE ITEM, BIOLOGICAL OR SELECT AGENT OUTSIDE THE U.S?

☒ **No**
☐ **Yes**

Section 2: Justification

2.1 DESCRIBE THE BACKGROUND, PURPOSE, AND ANTICIPATED FINDINGS OF THIS STUDY:

Vascular endothelial growth factor (VEGF) is one of the most potent inducers of angiogenesis known. VEGF induces its biologic functions through two main cell-surface receptors, Flt1 and Flk1. In addition to the full length transmembrane form of Flt1, the Flt1 gene also encodes a secreted, truncated form of the receptor via an mRNA in which a portion of intron 13 is preserved. This soluble Flt1 (sFlt1) retains high binding affinity for VEGF, and functions to negatively regulates the angiogenic activity of VEGF. The sequence of intron 13 appears to contain important regulatory regions that may be involved in sFlt1 mRNA formation. The purpose of this research iss to examine the sequence of human intron 13 and the adjacent exons (exons 12 and 13) for the presence of single nucleotide polymorphisms (SNPs) that may influence sFlt1 expression. It is especially of significance to look for SNPs at sites that have previously been identified as potential signals governing splicing or cleavage-polyadenylation in intron 13. Using the data found in the SNPbrowser Software (v3.5.2, Applied Biosystems) and NCBI SNP Database, a map of the region of interest in the Flt-1 gene was constructed with 23 SNPs, 2 of which were found in exons. However, SNP data in the Flt1 gene area is incomplete, and what data there is available is not validated. Finding, validating and subsequent studying of SNPs in this region is essential in determining whether individual Flt1 genotypes may influence the formation of one form of the Flt-1 receptor at the expense of the other, and therefore responsiveness to VEGF. Therefore, additional screening of this region with new DNA samples is essential in discovery of new SNPs and/or validating current SNP data.

In this study, we aim to do so by screening DNA from 100 unrelated, randomly chosen individuals. We will collect blood samples, extract the DNA from these blood samples and screen the SNPs in our region of interest.

2.2 EXPLAIN WHAT THE RESEARCH TEAM PLANS TO DO WITH THE STUDY RESULTS:
For example - publish or use for dissertation

Publish and use for PhD Dissertation

2.3 DESCRIBE THE RELATIONSHIP (IF ANY) THIS IRB APPLICATION HAS WITH ANY PREVIOUS OR UPCOMING VT IRB APPLICATIONS, INCLUDING INTERIM APPROVALS:

N/A

Section 3: Recruitment

3.1 DESCRIBE THE SUBJECT POOL, INCLUDING INCLUSION AND EXCLUSION CRITERIA AND NUMBER OF SUBJECTS:
Examples of inclusion/exclusion criteria - gender, age, health status, ethnicity

100 random individuals, male or female, above the age of 18, healthy and nonpregnant.

3.2 WILL EXISTING RECORDS BE USED TO IDENTIFY AND CONTACT / RECRUIT SUBJECTS?
Examples of existing records - directories, class roster, university records, educational records

☒ **No,** go to question 3.3
☐ **Yes,** answer questions within table

IF YES
Are these records private or public? ☐ Public ☐ Private, describe the researcher's privilege to the records:
Will student, faculty, and/or staff records or contact information be requested from the University? ☐ No ☐ Yes, visit the following link for further information: http://www.policies.vt.edu/index.php (policy no. 2010)

3.3 DESCRIBE RECRUITMENT METHODS, INCLUDING HOW THE STUDY WILL BE ADVERTISED OR INTRODUCED TO SUBJECTS:
[Include all recruitment materials with this application *(required for all protocols if data will be collected from people)* ***e.g., flyers/posters, invitation letter/e-mail, telephone recruitment script, SONA announcement, etc.]***

Recruitment will be via listservs and flyers on boards posted across campus. Content of posting as follows: **Volunteers needed to give blood samples!** **Researchers in the Biomedical Sciences Department at the Virginia Tech Vet School are currently studying the DNA sequence of the Flt1 gene. This gene is involved in angiogenesis (the formation of new blood vessels), a process that occurs in the function of a normal, healthy body (such as in wound healing) as well as in a diseased state (such as in the growth of a tumor in cancer).** **We need volunteers who are willing to provide a small blood sample in order to study the natural occurrence of differences in this gene. You must be 18 years or older (must show Hokie passport or valid State ID), healthy, nonpregnant and willing to sign a consent form. Your participation will be completely annonymous, with no documentation of any kind to link your blood sample to your identity.**

If you are willing to provide us with a blood sample, or would like more information, please contact Burouj Ajlouni at bajlouni@vt.edu or 540-231-9655

3.4 PROVIDE AN EXPLANATION FOR CHOOSING THIS POPULATION:

Note: the IRB must ensure that the risks and benefits of participating in a study are distributed equitably among the general population and that a specific population is not targeted because of ease of recruitment.

No specific population is targeted

Section 4: Consent Process

For more information about consent process and consent forms visit the following link: http://www.irb.vt.edu/pages/consent.htm

If feasible, researchers are advised and may be required to obtain signed consent from each participant unless obtaining signatures leads to an increase of risk (e.g., the only record linking the subject and the research would be the consent document and the principal risk would be potential harm resulting in a breach of confidentiality). Signed consent is typically not required for low risk questionnaires (consent is implied) unless audio/video recording or an in-person interview is involved. If researchers will not be obtaining signed consent, participants must, in most cases, be supplied with consent information in a different format (e.g., in recruitment document, at the beginning of survey instrument, read to participant over the phone, information sheet physically or verbally provided to participant).

4.1 CHECK ALL OF THE FOLLOWING THAT APPLY TO THIS STUDY'S CONSENT PROCESS:

☐ Verbal consent will be obtained from participants ***[include verbal script with this application]***
☒ Written/signed consent will be obtained from participants ***[include consent form(s) with this application]***
☐ Consent will be implied from the return of completed questionnaire. Note: The IRB recommends providing consent information in a recruitment document or at the beginning of the questionnaire (if the study only involves implied consent, skip to Section 5 below)
☐ Other, describe:

4.2 PROVIDE A GENERAL DESCRIPTION OF THE PROCESS THE RESEARCH TEAM WILL USE TO OBTAIN AND MAINTAIN INFORMED CONSENT:

Co-Investigator will obtain informed consent from participants. The trained helath professional who is responsible for taking the blood sample and labelling it will provide the participant with the consent form. The form must be signed before the blood sample will be extracted.

4.3 WHO, FROM THE RESEARCH TEAM, WILL BE OVERSEEING THE PROCESS AND OBTAINING CONSENT FROM SUBJECTS?

Co-Investigator Burouj Ajlouni

4.4 WHERE WILL THE CONSENT PROCESS TAKE PLACE?

At Schiffert Health Center, before the blood sample is taken

4.5 DURING WHAT POINT IN THE STUDY PROCESS WILL CONSENTING OCCUR?

Note: unless waived by the IRB, participants must be consented before completing any study procedure, including screening questionnaires.

Consent will take place before any part of the study begins, i.e. before a blood sample is collected or analyzed.

4.6 IF APPLICABLE, DESCRIBE HOW THE RESEARCHERS WILL GIVE SUBJECTS AMPLE TIME TO REVIEW THE CONSENT DOCUMENT BEFORE SIGNING:

Note: typically applicable for complex studies, studies involving more than one session, or studies involving more of a risk to subjects.

☒ Not applicable

Section 5: Procedures

5.1 PROVIDE A STEP-BY-STEP THOROUGH EXPLANATION OF ALL STUDY PROCEDURES EXPECTED FROM STUDY PARTICIPANTS, INCLUDING TIME COMMITMENT & LOCATION:

Location: Schiffert Health Center on the Virginia Tech Campus
Time Commitment: A maximum of 30-60 minutes will be required from test subjects. This includes time to go through informed consent process, to individually review the consent forms, sign (if agreeing to participate) and volunteer a blood sample. Only this one sample will be required, and no further participation or commitment will be required of any volunteers.

5.2 DESCRIBE HOW DATA WILL BE COLLECTED AND RECORDED:
[Include all data documents (e.g., questionnaire, interview questions, etc.) with protocol]

Blood will be collected at Schiffert Health Center via venipuncture. About 10mL (2 teaspoons) of blood will be collected from each paticipant. Once collection process is done, co-Investigator will collect blood samples which have been labelled with an abstract number and proceed to extract DNA from these blood samples. These DNA samples will be stored in the laboratory of the PI (Dr. William Huckle).

5.3 DOES THE PROJECT INVOLVE ONLINE RESEARCH ACTIVITES (INCLUDES ENROLLMENT, RECRUITMENT, SURVEYS)?
View the "Policy for Online Research Data Collection Activities Involving Human Subjects" at http://www.irb.vt.edu/documents/onlinepolicy.pdf

☒ **No,** go to question 6.1
☐ **Yes,** answer questions within table →

IF YES
Identify the service / program that will be used: ☐ www.survey.vt.edu, go to question 6.1 ☐ Blackboard, go to question 6.1 ☐ Center for Survey Research, go to question 6.1 ☐ Other **IF OTHER:** Name of service / program: URL: This service is… ☐ Included on the list found at: http://www.irb.vt.edu/pages/validated.htm ☐ Approved by VT IT Security ***[Include approval correspondence with protocol]*** ☐ An external service with proper SSL or similar encryption (https://) on the login (if applicable) and all other data collection pages. ☐ None of the above (note: only permissible if this is a collaborative project in which VT individuals are only responsible for data analysis, consulting, or recruitment)

Section 6: Risks and Benefits

6.1 WHAT ARE THE POTENTIAL RISKS (E.G., EMOTIONAL, PHYSICAL, SOCIAL, LEGAL, ECONOMIC, OR DIGNITY) TO STUDY PARTICIPANTS?

There are very minimal risks associated with giving a blood sample.
Some risks (rare, but possible) may include:
- Excessive bleeding
- Fainting or feeling light-headed
- Hematoma (blood accumulating under the skin)
- Infection (a slight risk any time the skin is broken)

6.2 EXPLAIN THE STUDY'S EFFORTS TO REDUCE POTENTIAL RISKS TO SUBJECTS:

- A trained health professional will be collecting the blood sample and will follow all health and safety regulations, such as maintaining aseptic conditions during the collection process.
- Skin will be swabbed with alcohol before skin is punctured, to minize risk of infection.
- Only 10mL of blood will be collected (very small amount, thus low risk).
- Those considered "high risk" (bleeding disorders, compromised immune system, on anticoagulant medication) will be excluded from participation.
- Volunteers must indicate that they have had a meal within four hours prior to volunteering a blood sample.

6.3 WHAT ARE THE DIRECT OR INDIRECT ANTICIPATED BENEFITS TO STUDY PARTICIPANTS AND/OR SOCIETY?

There is no direct benefit to participants. However, the benefit is to the science at hand and to society in general. This study will allow us to validate the prescence of SNPs present in the Flt1 gene area. SNPs have been found to play a role in many major diseases, and SNPs in the Flt1 gene may play a role in the development of angiogenesis related diseases such as cancer and diabetes. The results of this study will help us in understanding the function of the Flt1 gene and how SNPs affect this function. This increased understanding will in turn allow for a broader understanding of the role of Flt1 in the development and/or progression of disease, which may in turn lead to the design of newer and more effective therapies.

Section 7: Full Board Assessment

7.1 DOES THE RESEARCH INVOLVE MICROWAVES/X-RAYS, OR GENERAL ANESTHESIA OR SEDATION?

☒ **No**
☐ **Yes**

7.2 DO RESEARCH ACTIVITIES INVOLVE PRISONERS, PREGNANT WOMEN, FETUSES, HUMAN IN VITRO FERTILIZATION, OR MENTALLY DISABLED PERSONS?

☒ **No,** go to question 7.3
☐ **Yes,** answer questions within table

IF YES
This research involves: ☐ Prisoners ☐ Pregnant women ☐ Fetuses ☐ Human in vitro fertilization ☐ Mentally disabled persons

7.3 DOES THIS STUDY INVOLVE MORE THAN MINIMAL RISK TO STUDY PARTICIPANTS?

Minimal risk means that the probability and magnitude of harm or discomfort anticipated in the research are not greater in and of themselves than those ordinarily encountered in daily activities or during the performance of routine physical or psychological examinations or tests. Examples of research involving greater than minimal risk include collecting data about abuse or illegal activities. Note: if the project qualifies for Exempt review (http://www.irb.vt.edu/pages/categories.htm), it will not need to go to the Full Board.

☒ **No**
☐ **Yes**

IF YOU ANSWERED "YES" TO ***ANY ONE*** OF THE ABOVE QUESTIONS, 7.1, 7.2, OR 7.3, THE BOARD MAY REVIEW THE PROJECT'S APPLICATION MATERIALS AT ITS MONTHLY MEETING. VIEW THE FOLLOWING LINK FOR DEADLINES AND ADDITIONAL INFORMATION: http://www.irb.vt.edu/pages/deadlines.htm

Section 8: Confidentiality / Anonymity

For more information about confidentiality and anonymity visit the following link: http://www.irb.vt.edu/pages/confidentiality.htm

8.1 WILL PERSONALLY IDENTIFYING STUDY RESULTS OR DATA BE RELEASED TO ANYONE OUTSIDE OF THE RESEARCH TEAM?

For example – to the funding agency or outside data analyst, or participants identified in publications with individual consent

☒ **No**
☐ **Yes,** to whom will identifying data be released?

8.2 WILL ANY STUDY FILES CONTAIN PARTICIPANT IDENTIFYING INFORMATION (E.G., NAME, CONTACT INFORMATION, VIDEO/AUDIO RECORDINGS)?

Note: if collecting signatures on a consent form, select "Yes."

☐ **No,** go to question 8.3
☒ **Yes,** answer questions within table →

IF YES
Describe if/how the study will utilize study codes: All samples will be given an abstract number (1,2,3 etc) that will NOT be linked to the consent forms. Therefore, samples will be completely annonymous.
If applicable, where will the key [i.e., linked code and identifying information document (for instance, John Doe = study ID 001)] **be stored and who will have access? No key will be made as samples will not be linked to consent forms.** *Note: the key should be stored separately from subjects' completed data documents and accessibility should be limited.*
The IRB strongly suggests and may require that all data documents (e.g., questionnaire responses, interview responses, etc.) do not include or request identifying information (e.g., name, contact information, etc.) from participants. If you need to link subjects' identifying information to subjects' data documents, use a study ID/code on all data documents.

8.3 WHERE WILL DATA BE STORED?

Examples of data - questionnaire, interview responses, downloaded online survey data, observation recordings, biological samples

In the lab of the PI, in the form of DNA samples. Data in form of experimental results will be stored in PI's office. Both sites are routinely locked when unoccupied.

8.4 WHO WILL HAVE ACCESS TO STUDY DATA?

PI and co-investigator.

8.5 DESCRIBE THE PLANS FOR RETAINING OR DESTROYING THE STUDY DATA

Consent forms will be stored securely for 3 years as per federal regulations, but there will be no identiable information that will link identity to the sample. They will be destroyed appropriately after three years have passed.
Samples that were collected will be analyzed and results will be published.

8.6 DOES THIS STUDY REQUEST INFORMATION FROM PARTICIPANTS REGARDING ILLEGAL BEHAVIOR?

☒ **No,** go to question 9.1
☐ **Yes,** answer questions within table

IF YES
Does the study plan to obtain a Certificate of Confidentiality? ☐ No ☐ Yes (Note: participants must be fully informed of the conditions of the Certificate of Confidentiality within the consent process and form) *For more information about Certificates of Confidentiality, visit the following link:* *http://www.irb.vt.edu/pages/coc.htm*

Section 9: Compensation

For more information about compensating subjects, visit the following link: http://www.irb.vt.edu/pages/compensation.htm

9.1 WILL SUBJECTS BE COMPENSATED FOR THEIR PARTICIPATION?

☒ **No,** go to question 10.1
☐ **Yes,** answer questions within table

IF YES
What is the amount of compensation?
Will compensation be prorated? ☐ Yes, please describe: ☐ No, explain why and clarify whether subjects will receive full compensation if they withdraw from the study? *Unless justified by the researcher, compensation should be prorated based on duration of study participation. Payment must not be contingent upon completion of study procedures. In other words, even if the subject decides to withdraw from the study, he/she should be compensated, at least partially, based on what study procedures he/she has completed.*

Section 10: Audio / Video Recording

For more information about audio/video recording participants, visit the following link: http://www.irb.vt.edu/pages/recordings.htm

10.1 WILL YOUR STUDY INVOLVE VIDEO AND/OR AUDIO RECORDING?

☒ **No,** go to question 11.1
☐ **Yes,** answer questions within table

IF YES
This project involves: ☐ Audio recordings only ☐ Video recordings only ☐ Both video and audio recordings
Provide compelling justification for the use of audio/video recording:
How will data within the recordings be retrieved / transcribed?
How and where will recordings (e.g., tapes, digital data, data backups) be stored to ensure security?
Who will have access to the recordings?
Who will transcribe the recordings?
When will the recordings be erased / destroyed?

Section 11: Research Involving Students

11.1 DOES THIS PROJECT INCLUDE STUDENTS AS PARTICIPANTS?

☐ **No,** go to question 12.1
☒ **Yes,** answer questions within table

IF YES
Does this study involve conducting research with students of the researcher? ☒ No ☐ Yes, describe safeguards the study will implement to protect against coercion or undue influence for participation: *Note: if it is feasible to use students from a class of students not under the instruction of the researcher, the IRB recommends and may require doing so.*
Will the study need to access student records (e.g., SAT, GPA, or GRE scores)? ☒ No ☐ Yes ***[Include release consent / assent / permission form(s) with this application]***

11.2 DOES THIS PROJECT INCLUDE ELEMENTARY, JUNIOR, OR HIGH SCHOOL STUDENTS?

☒ **No,** go to question 11.3
☐ **Yes,** answer questions within table

IF YES
Will study procedures be completed during school hours? ☐ No ☐ Yes **If yes,** **Students not included in the study may view other students' involvement with the research during school time as unfair. Address this issue and how the study will reduce this outcome:** **Missing out on regular class time or seeing other students participate may influence a student's decision to participate. Address how the study will reduce this outcome:**
Is the school's approval letter(s) attached to this submission? ☐ Yes ☐ No, project involves Montgomery County Public Schools (MCPS) ☐ No, explain why: *You will need to obtain school approval (if involving MCPS, click here: http://www.irb.vt.edu/pages/mcps.htm). Approval is typically granted by the superintendent, principal, and classroom teacher (in that order). Approval by an individual teacher is insufficient. School approval, in the form of a letter or a memorandum should accompany the approval request to the IRB.*

11.3 DOES THIS PROJECT INCLUDE COLLEGE STUDENTS?

☐ **No,** go to question 12.1
☒ **Yes,** answer questions within table

IF YES
Some college students might be minors. Indicate whether these minors will be included in the research or actively excluded: ☐ Included ☒ Actively excluded, describe how the study will ensure that minors will not be included: **Must show Hokie passport or State ID before given a consent form. Also, signing of the consent form will have a statement indicating that the person signing must be 18 years of age or older.**
Will extra credit be offered to subjects? ☒ No ☐ Yes **If yes,** **What will be offered to subjects as an equal alternative to receiving extra credit without participating in this study?** **Include a description of the extra credit (e.g., amount) to be provided within question 9.1 ("IF YES" table)**

Section 12: Research Involving Minors

12.1 DOES THIS PROJECT INVOLVE MINORS (UNDER THE AGE OF 18 IN VIRGINIA)?

Note: age constituting a minor may differ in other States.

☒ **No,** go to question 13.1
☐ **Yes,** answer questions within table

IF YES
Does the project reasonably pose a risk of reports of current threats of abuse and/or suicide? ☐ No ☐ Yes, thoroughly explain how the study will react to such reports: *Note: subjects and parents must be fully informed of the fact that researchers must report threats of suicide or suspected/reported abuse to the appropriate authorities within the Confidentiality section of the Consent, Assent, and/or Permission documents.*
Are you requesting a waiver of parental permission (i.e., parent uninformed of child's involvement)? ☐ No, **both** parents/guardians will provide their permission, if possible. ☐ No, **only one** parent/guardian will provide permission. ☐ Yes, describe below how your research meets **all** of the following criteria (A-D): Criteria A - The research involves no more than minimal risk to the subjects: Criteria B - The waiver will not adversely affect the rights and welfare of the subjects: Criteria C - The research could not practicably be carried out without the waiver: Criteria D - (Optional) Parents will be provided with additional pertinent information after participation:
Is it possible that minor research participants will reach the legal age of consent (18 in Virginia) while enrolled in this study? ☐ No ☐ Yes, will the investigators seek and obtain the legally effective informed consent (in place of the minors' previously provided assent and parents' permission) for the now-adult subjects for any ongoing interactions with the subjects, or analysis of subjects' data? If yes, explain how: *For more information about minors reaching legal age during enrollment, visit the following link:* *http://www.irb.vt.edu/pages/assent.htm*
*The procedure for obtaining assent from minors and permission from the minor's guardian(s) must be described in **Section 4** (Consent Process) of this form.*

Section 13: Research Involving Deception

For more information about involving deception in research and for assistance with developing your debriefing form, visit our website at http://www.irb.vt.edu/pages/deception.htm

13.1 DOES THIS PROJECT INVOLVE DECEPTION?

☒ **No,** go to question 14.1
☐ **Yes,** answer questions within table

IF YES
Describe the deception:
Why is the use of deception necessary for this project?
Describe the debriefing process:

[Include debriefing form with this application]
Provide an explanation of how the study meets <u>all</u> the following criteria (A-D) for an alteration of consent: Criteria A - The research involves no more than minimal risk to the subjects: Criteria B - The alteration will not adversely affect the rights and welfare of the subjects: Criteria C - The research could not practicably be carried out without the alteration: Criteria D - (Optional) Subjects will be provided with additional pertinent information after participation (i.e., debriefing for studies involving deception): *By nature, studies involving deception cannot provide subjects with a complete description of the study during the consent process; therefore, the IRB must allow (by granting an alteration of consent) a consent process which does not include, or which alters, some or all of the elements of informed consent.*
The IRB requests that the researcher use the title "Information Sheet" instead of "Consent Form" on the document used to obtain subjects' signatures to participate in the research. This will adequately reflect the fact that the subject cannot fully consent to the research without the researcher fully disclosing the true intent of the research.

Section 14: Research Involving Existing Data

14.1 WILL THIS PROJECT INVOLVE THE COLLECTION OR STUDY/ANALYSIS OF EXISTING DATA DOCUMENTS, RECORDS, PATHOLOGICAL SPECIMENS, OR DIAGNOSTIC SPECIMENS?

Please note: it is not considered existing data if a researcher transfers to Virginia Tech from another institution and will be conducting data analysis of an on-going study.

☒ **No,** you are finished with the application
☐ **Yes,** answer questions within table

IF YES
From where does the existing data originate?
Provide a detailed description of the existing data that will be collected or studied/analyzed:
Is the source of the data public? ☐ No, continue with the next question ☐ Yes, you are finished with this application
Will any individual associated with this project (internal or external) have access to or be provided with existing data containing information which would enable the identification of subjects: ▪ **Directly** (e.g., by name, phone number, address, email address, social security number, student ID number), or ▪ **Indirectly through study codes** even if the researcher or research team does not have access to the master list linking study codes to identifiable information such as name, student ID number, etc or ▪ **Indirectly through the use of information that could reasonably be used in combination to identify an individual** (e.g., demographics) ☐ No, collected/analyzed data will be completely de-identified ☐ Yes, **If yes,** *Research will not qualify for exempt review; therefore, if feasible, written consent must be obtained*

from individuals whose data will be collected / analyzed, unless this requirement is waived by the IRB.

Will written/signed or verbal consent be obtained from participants prior to the analysis of collected data? -select one-

This research protocol represents a contract between all research personnel associated with the project, the University, and federal government; therefore, must be followed accordingly and kept current.

Proposed modifications must be approved by the IRB prior to implementation except where necessary to eliminate apparent immediate hazards to the human subjects.

Do not begin human subjects activities until you receive an IRB approval letter via email.

----------END----------

2. IRB Consent Form

VIRGINIA POLYTECHNIC INSTITUTE AND STATE UNIVERSITY

Consent Form: "Screening Flt1 Gene for SNPs"

Principal Investigator: Dr. W.R. Huckle
Co-Investigator: Burouj Ajlouni, PhD Candidate

I. Purpose of this Research/Project

We are interested in studying the DNA sequence of the Flt1 gene, which codes for several proteins involved in angiogenesis (the formation of new blood vessels). Angiogenesis is a process that occurs in the function of a normal, healthy body (such as in wound healing) as well as in a diseased state (such as in the growth of a tumor in cancer).

We would like to study single base pair changes in the DNA, called single nucleotide polymorphisms (or SNPs) which may have an effect on how this Flt1 gene functions. A better understanding of this gene and the function of its proteins will allow us to better understand how the process of angiogenesis work and may potentially have a benefit in researching new treatment protocols for certain diseases such as cancer.

We would like to collect 100 blood samples from 100 unrelated, randomly chosen healthy individuals. This blood will be used to extract DNA upon which we will study the SNPs in the Flt1 gene.

Because SNPs vary by ethnic pools, we would require volunteers to indicate their racial status. Race will not be used to exclude nor encourage any volunteers over others – it will simply be used by the researchers to classify the sample.

II. Role of Volunteer

As a volunteer, you will be asked a few questions about the following:
a) Your age: you must be 18 years or older to participate
b) Your health status: you must not be pregnant (females), you must not have any bleeding disorders or a history of any bleeding disorders that may increase the risk of side effects when giving a blood sample and you must also not have a compromised immune system.
c) Medications taken in the past 12 hours: You must not have taken any kind of thrombolytic medications that would slow blood clotting (including "aspiring") within the past 12 hours.
d) Your last meal: You must have eaten a meal within the 4 hours prior to giving your blood sample.
e) Your race/ethnic background: because we are studying SNPs, which vary by

ethnic background, it is important for us to know your background as precisely as possibly.
Once you have answered these questions, you will be asked to donate a small blood sample. This sample will never be used to identify you as an individual, nor will you be contacted at a further date to provide anything else.

III. Risk

There are very minimal risks associated with giving a blood sample.
Some risks (rare, but possible) may include:
• Bruising
• Excessive bleeding
• Fainting or feeling light-headed
• Hematoma (blood accumulating under the skin)
• Infection (a slight risk any time the skin is broken)

In order to minimize risks, the following will be done:
• Blood sample will be collected by a trained health professional at Schiffert Health Center, and all health and safety regulations will be followed, such as maintaining aseptic conditions during the collection process.
• Skin will be swabbed with alcohol before skin is punctured, to minimize risk of infection.
• Only 10mL of blood will be collected (very small amount, thus low risk).
• Those considered "high risk" (bleeding disorders, compromised immune system, on anticoagulant medication) will be excluded from participation.
• Volunteers must indicate that they have had a meal within four hours prior to volunteering a blood sample, to avoid fainting/light-headedness.

IV. Benefits

There is no direct benefit to you, as a volunteer. However, the benefit is to the science at hand and to society in general. This study will allow us to validate the presence of SNPs present in the Flt1 gene area. SNPs have been found to play a role in many major diseases, and SNPs in the Flt1 gene may play a role in the development of angiogenesis related diseases such as cancer and diabetes. The results of this study will help us in understanding the function of the Flt1 gene and how SNPs affect this function. This increased understanding will in turn allow for a broader understanding of the role of Flt1 in the development and/or progression of disease, which may in turn lead to the design of newer and more effective therapies.

V. Extent of Anonymity and Confidentiality

Complete anonymity is promised. Your only link to this study will be a consent form, which will NOT be linked in any way to your blood sample.

Your blood sample will be given an abstract number that cannot be identified by anyone, not even the investigators. Once your sample is collected, it will be stored amongst all the other collected samples and become virtually unidentifiable with no record of any link to you, the volunteer.

The only link you will have to this project is your signed consent form, which will be stored securely with the Primary Investigator, Dr. William Huckle. Consent forms will be stored in a secure and confidential manner for three years, as required by Federal laws. After three years have passed, these consent forms will be destroyed and there will be no other evidence linking you to this project.

Also, please note that there will be no record of your participation in this study in your Schiffert Health file nor in any insurance documents. This study will ensure complete confidentiality of your participation (no one besides the research group will know of your participation in this study), and complete anonymity of your sample (any given blood sample is in no way identifiable to the person who volunteered it).

Under no conditions will your anonymity be broken and at no time will the researchers release any identifiable information to anyone other than individuals working on the project. However, it is possible that the Institutional Review Board (IRB) may view this study's collected data for auditing purposes. The IRB is responsible for the oversight of the protection of human subjects involved in research.

VI. Compensation

There will be no compensation to any volunteers for participating in this study.

VII. Subject's Responsibilities

I voluntarily agree to participate in this study. I have the following responsibilities:

I agree to show my ID to the person collecting my blood sample, in order to verify that I am 18 years of age or older.
By signing below, I am affirming that <u>none of the following is applicable to me</u>:
a) I am pregnant (females), I suffer from bleeding disorders or have a history of any bleeding disorders that may increase the risk of side effects when giving a blood sample

b) I have a compromised immune system.
a) I have taken any kind of medications that would slow blood clotting (including "aspirin") within the past 12 hours.
b) I have not eaten a meal within the 4 hours prior to giving a blood sample.

I also agree to disclose my ethnic background: (please circle one; if mixed race, please choose "other" and explain):
Caucasian _____
African American _____
Japanese _____
Chinese _____
Other (please explain): _____

VIII. Subject's Permission

I have read the Consent Form and conditions of this project. I have had all my questions answered. No promises or guarantee of benefits have been made to encourage me to participate.

I hereby acknowledge the above and give my voluntary consent:

_____ Date_____
Subject signature
_____ Date _____
Witness (Optional except for certain classes of subjects)

Should I have any pertinent questions about this research or its conduct, and research subjects' rights, and whom to contact in the event of a research-related injury to the subject, I may contact:

Burouj Ajlouni, 540-231-9655, bajlouni@vt.edu, (Investigator)

Dr. W. Huckle, 540-231-9655, wrhuckle@vt.edu (Faculty Advisor/Primary Investigator)

If I should have any questions about the protection of human research participants regarding this study, I may contact Dr. David Moore, Chair Virginia Tech Institutional Review Board for the Protection of Human Subjects
Telephone: (540) 231-4991
Email: moored@vt.edu
Address: Office of Research Compliance, 2000 Kraft Drive, Suite 2000 (0497), Blacksburg, VA 24060.

3. IRB Approval Letter

VirginiaTech

Office of Research Compliance
Institutional Review Board
2000 Kraft Drive, Suite 2000 (0497)
Blacksburg, Virginia 24061
540/231-4991 Fax 540/231-0959
e-mail moored@vt.edu
www.irb.vt.edu
FWA00000572(expires 1/20/2010)
IRB # is IRB00000667

DATE: May 14, 2009

MEMORANDUM

TO: William R. Huckle
Burouj Ajlouni

Approval date: 10/21/2008
Continuing Review Due Date:10/6/2009
Expiration Date: 10/20/2009

FROM: David M. Moore

SUBJECT: **IRB Amendment 2 Approval:** "Screening Flt1 Region of Interest for SNPs", OSP #445076, IRB # 08-636

This memo is regarding the above referenced protocol which was previously granted approval by the IRB on October 21, 2008. You subsequently requested permission to amend your IRB application. Since the requested amendment is nonsubstantive in nature, I, as Chair of the Virginia Tech Institutional Review Board, have granted approval for requested protocol amendment, effective as of May 14, 2009. The anniversary date will remain the same as the original approval date.

As an investigator of human subjects, your responsibilities include the following:

1. Report promptly proposed changes in previously approved human subject research activities to the IRB, including changes to your study forms, procedures and investigators, regardless of how minor. The proposed changes must not be initiated without IRB review and approval, except where necessary to eliminate apparent immediate hazards to the subjects.
2. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.
3. Report promptly to the IRB of the study's closing (i.e., data collecting and data analysis complete at Virginia Tech). If the study is to continue past the expiration date (listed above), investigators must submit a request for continuing review prior to the continuing review due date (listed above). It is the researcher's responsibility to obtain re-approval from the IRB before the study's expiration date.
4. If re-approval is not obtained (unless the study has been reported to the IRB as closed) prior to the expiration date, all activities involving human subjects and data analysis must cease immediately, except where necessary to eliminate apparent immediate hazards to the subjects.

cc: File

4. IRB Amendment Request #1

VirginiaTech

Institutional Review Board
Amendment Request Form

INSTRUCTIONS: Email completed form and all revised and/or new study documents to irb@vt.edu (PDFs preferred).

Note: The project's IRB-approved Research Protocol (previously entitled Initial Review Application) must be kept current and followed throughout the life of the project. It is advised that it be reviewed prior to the submission of an amendment request to ensure all changes are reflected. All study documents are subject to audit.

1. IRB NUMBER:

08-636

2. PROJECT TITLE:

Screening Flt1 Region of Interest for SNPs

3. PRINCIPAL INVESTIGATOR

Name:	Email address:
William Huckle	**wrhuckle@vt.edu**

4. REQUESTING AMENDMENT TO:

- ☐ Research Protocol (or Initial Review Application)
- ☐ Consent form
- ☒ Recruitment materials
- ☐ Data document (e.g., survey instrument, interview questions)
- ☐ Research personnel
- ☐ Other

5. DESCRIBE THE AMENDMENT BEING REQUESTED:

Note: with each requested change, provide a detailed description of where within the study documents (e.g., Research Protocol, survey instrument, etc.) the changes are reflected (e.g., page number, question #, etc.)

In addition to the already approved listserv listing in the original IRB form, we would also like to use a flyer to recruit volunteers. Please see attached flyer for format.

6. HAVE THESE REQUESTED CHANGES BEEN INITIATED?

- ☒ No
- ☐ Yes, why were these changes initiated prior to being approved (see bottom of page)?

7. HOW WILL THE PROPOSED AMENDMENT AFFECT STUDY PARTICIPANTS?

There will be no change in the study; only the use of a flyer to recruit volunteers

Federal regulations require IRB approval prior to changing a research procedure or deviating from IRB-approved documents unless it is in the best interest of or for the safety of study participants.

5. IRB Amendment Approval Letter #1

VirginiaTech

Office of Research Compliance
Institutional Review Board
2000 Kraft Drive, Suite 2000 (0497)
Blacksburg, Virginia 24061
540/231-4991 Fax 540/231-0959
e-mail moored@vt.edu
www.irb.vt.edu
FWA00000572(expires 1/20/2010)
IRB # is IRB00000667

DATE: April 6, 2009

MEMORANDUM

TO: William R. Huckle
Burouj Ajlouni

Approval date: 10/21/2008
Continuing Review Due Date:10/6/2009
Expiration Date: 10/20/2009

FROM: David M. Moore

SUBJECT: **IRB Amendment 1 Approval:** "Screening Flt1 Region of Interest for SNPs", OSP #445076, IRB # 08-636

This memo is regarding the above referenced protocol which was previously granted approval by the IRB on October 21, 2008. You subsequently requested permission to amend your IRB application. Since the requested amendment is nonsubstantive in nature, I, as Chair of the Virginia Tech Institutional Review Board, have granted approval for requested protocol amendment, effective as of April 6, 2009. The anniversary date will remain the same as the original approval date.

As an investigator of human subjects, your responsibilities include the following:

1. Report promptly proposed changes in previously approved human subject research activities to the IRB, including changes to your study forms, procedures and investigators, regardless of how minor. The proposed changes must not be initiated without IRB review and approval, except where necessary to eliminate apparent immediate hazards to the subjects.
2. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.
3. Report promptly to the IRB of the study's closing (i.e., data collecting and data analysis complete at Virginia Tech). If the study is to continue past the expiration date (listed above), investigators must submit a request for continuing review prior to the continuing review due date (listed above). It is the researcher's responsibility to obtain re-approval from the IRB before the study's expiration date.
4. If re-approval is not obtained (unless the study has been reported to the IRB as closed) prior to the expiration date, all activities involving human subjects and data analysis must cease immediately, except where necessary to eliminate apparent immediate hazards to the subjects.

cc: File

6. IRB Amendment #2

VirginiaTech

Institutional Review Board
Amendment Request Form

INSTRUCTIONS: Email completed form and all revised and/or new study documents to irb@vt.edu (PDFs preferred).

Note: The project's IRB-approved Research Protocol (previously entitled Initial Review Application) must be kept current and followed throughout the life of the project. It is advised that it be reviewed prior to the submission of an amendment request to ensure all changes are reflected. All study documents are subject to audit.

1. IRB NUMBER:

08-636

2. PROJECT TITLE:

Screening Flt1 Region of Interest for SNPs

3. PRINCIPAL INVESTIGATOR

Name:	Email address:
William Huckle	**wrhuckle@vt.edu**

4. REQUESTING AMENDMENT TO:

- ☒ Research Protocol (or Initial Review Application)
- ☐ Consent form
- ☐ Recruitment materials
- ☐ Data document (e.g., survey instrument, interview questions)
- ☐ Research personnel
- ☐ Other

5. DESCRIBE THE AMENDMENT BEING REQUESTED:

Note: with each requested change, provide a detailed description of where within the study documents (e.g., Research Protocol, survey instrument, etc.) the changes are reflected (e.g., page number, question #, etc.)

On pages 5 and 6 of the Research Protocol, we describe the informed consent process and blood collection. In addition to the process described there, we would also like to add the following:
- Informed consent may also be obtained with a phone conversation with Investigator, with the signature on the consent form being collected by a third party on behalf of the investigator.
- In addition to blood collection at Schiffert Health Center, we would also like to collect blood at other sites, with blood collection being performed by a licensed phlebotomist who will follow all health and safety regulations, such as maintaining aseptic conditions during the collection process.

6. HAVE THESE REQUESTED CHANGES BEEN INITIATED?

- ☒ No
- ☐ Yes, why were these changes initiated prior to being approved (see bottom of page)?

7. HOW WILL THE PROPOSED AMENDMENT AFFECT STUDY PARTICIPANTS?

This change will not affect the study participants, only the location of where the blood collection takes place.

Federal regulations require IRB approval prior to changing a research procedure or deviating from IRB-approved documents unless it is in the best interest of or for the safety of study participants.

7. IRB Amendment Approval #2

VirginiaTech

Office of Research Compliance
Institutional Review Board
2000 Kraft Drive, Suite 2000 (0497)
Blacksburg, Virginia 24061
540/231-4991 Fax 540/231-0959
e-mail moored@vt.edu
www.irb.vt.edu
FWA00000572(expires 1/20/2010)
IRB # is IRB00000667

DATE: May 14, 2009

MEMORANDUM

TO: William R. Huckle
Burouj Ajlouni

Approval date: 10/21/2008
Continuing Review Due Date:10/6/2009
Expiration Date: 10/20/2009

FROM: David M. Moore

SUBJECT: **IRB Amendment 2 Approval:** "Screening Flt1 Region of Interest for SNPs", OSP #445076, IRB # 08-636

This memo is regarding the above referenced protocol which was previously granted approval by the IRB on October 21, 2008. You subsequently requested permission to amend your IRB application. Since the requested amendment is nonsubstantive in nature, I, as Chair of the Virginia Tech Institutional Review Board, have granted approval for requested protocol amendment, effective as of May 14, 2009. The anniversary date will remain the same as the original approval date.

As an investigator of human subjects, your responsibilities include the following:

1. Report promptly proposed changes in previously approved human subject research activities to the IRB, including changes to your study forms, procedures and investigators, regardless of how minor. The proposed changes must not be initiated without IRB review and approval, except where necessary to eliminate apparent immediate hazards to the subjects.
2. Report promptly to the IRB any injuries or other unanticipated or adverse events involving risks or harms to human research subjects or others.
3. Report promptly to the IRB of the study's closing (i.e., data collecting and data analysis complete at Virginia Tech). If the study is to continue past the expiration date (listed above), investigators must submit a request for continuing review prior to the continuing review due date (listed above). It is the researcher's responsibility to obtain re-approval from the IRB before the study's expiration date.
4. If re-approval is not obtained (unless the study has been reported to the IRB as closed) prior to the expiration date, all activities involving human subjects and data analysis must cease immediately, except where necessary to eliminate apparent immediate hazards to the subjects.

cc: File

8. Flyer used to recruit volunteers

NEEDED TO GIVE BLOOD SAMPLES FOR

CANCER RESEARCH

Researchers in the Biomedical Sciences Department at the Virginia Tech Vet School are currently studying the DNA sequence of the Flt1 gene. This gene is involved in angiogenesis (the formation of new blood vessels), a process that occurs in the function of a normal, healthy body (such as in wound healing) as well as in a diseased state (such as in the growth of a tumor in cancer).

We need volunteers who are willing to provide a small blood sample in order to study the natural occurrence of differences in this gene.

You must be 18 years or older (must show Hokie passport or valid State ID), healthy, nonpregnant and willing to sign a consent form.

Your participation will be completely annonymous, with no documentation of any kind to link your blood sample to your identity.

If you are willing to provide us with a blood sample, or would like more information, please contact Burouj Ajlouni at bajlouni@vt.edu or 540-538-0752 / 540-231-9655

Appendix D: PERL Code for The EXONerator (Chapter 4)

```
#!/usr/bin/perl  - w

#***************Hash of amino acid code, with aa 1 letter code as the key**********
my(%genetic_code) = (
          'S' => ['TCA', 'TCC','TCG','TCT', 'AGT', 'AGC'], #Serine
          'F' => ['TTC', 'TTT'], #Phenylalanine
          'L' => ['TTA', 'TTG', 'CTA','CTC','CTG','CTT'], #Leucine
          'Y' => ['TAC', 'TAT'], #Tyrosine
          'C' => ['TGC', 'TGT'], #Cysteine
          'W' => ['TGG'], #Tryptophan
          'P' => ['CCA', 'CCC', 'CCG', 'CCT'], #Proline
          'H' => ['CAT', 'CAC'], #Histidine
          'Q' => ['CAA', 'CAG'], #Glutamine
          'R' => ['CGA', 'CGC','CGG','CGT', 'AGA', 'AGG'], #Arginine
          'I' => ['ATA', 'ATC', 'ATT'], #Isoleucine
          'M' => ['ATG'], #Methionine
          'T' => ['ACA', 'ACC', 'ACG', 'ACT'], #Threonine
          'N' => ['AAC', 'AAT'], #Asparagine
          'K' => ['AAA', 'AAG'], #Lysine
          'V' => ['GTA', 'GTC', 'GTG', 'GTT'], #Valine
          'A' => ['GCA', 'GCC', 'GCG', 'GCT'], #Alanine
          'D' => ['GAC', 'GAT'], #Aspartic Acid
          'E' => ['GAA', 'GAG'], #Glutamic Acid
          'G' => ['GGA', 'GGC', 'GGG', 'GGT'], #Glycine
          '*STOP' => ['TAA', 'TAG', 'TGA'], #Stop Codon
     );

#********************3D ARRAYS FOR PROTEIN SCORES**********************************


#2D Array for SF2 scores with nt ordered ACGT:0123
my @SF2scores = (
[-1.14, 1.37, -0.21, -1.58], #Scores for position 1, nt ordered ACGT:0123
[0.62, -1.1, 0.17, -0.5], #Scores for position 2, nt ordered ACGT:0123
[-1.58, 0.73, 0.48, -1.58], #Scores for position 3, nt ordered ACGT:0123
[1.32, 0.33, -1.58, -1.13], #Scores for position 4, nt ordered ACGT:0123
[-1.58, 0.94, 0.33, -1.58], #Scores for position 5, nt ordered ACGT:0123
[-1.58, -1.58, 0.99, -1.13], #Scores for position 6, nt ordered ACGT:0123
[0.62, -1.58, -0.11, 0.27] #Scores for position 7, nt ordered ACGT:0123
);
#
#3D Array for SC35 scores with nt ordered ACGT:0123
my @SC35scores = (
[-0.88, -1.16, 0.87, -1.18], #Scores for position 1, nt ordered ACGT:0123
[0.09, -1.58, 0.45, -0.2], #Scores for position 2, nt ordered ACGT:0123
```

```
[-0.06, 0.95, -1.36, 0.38], #Scores for position 3, nt ordered ACGT:0123
[-1.58, 1.11, -1.58, 0.88], #Scores for position 4, nt ordered ACGT:0123
[0.09, 0.56, -0.33, -0.2], #Scores for position 5, nt ordered ACGT:0123
[-0.41, 0.86, -0.05, -0.86], #Scores for position 6, nt ordered ACGT:0123
[-0.06, 0.32, -1.36, 0.96], #Scores for position 7, nt ordered ACGT:0123
[0.23, -1.58, 0.68, -1.58] #Scores for position 8, nt ordered ACGT:0123
);
#
#3D Array for SRp40 scores with nt ordered ACGT:0123
my @SRp40scores = (
[-0.13, 0.56, -1.58, 0.92], #Scores for position 1, nt ordered ACGT:0123
[-1.58, 0.68, -0.14, 0.37], #Scores for position 2, nt ordered ACGT:0123
[1.28, -1.12, -1.33, 0.23], #Scores for position 3, nt ordered ACGT:0123
[-0.33, 1.24, -0.48, -1.14], #Scores for position 4, nt ordered ACGT:0123
[0.97, -0.77, -1.58, 0.72], #Scores for position 5, nt ordered ACGT:0123
[-0.13, 0.13, 0.44, -1.58], #Scores for position 6, nt ordered ACGT:0123
[-1.58, -0.05, 0.8, -1.58] #Scores for position 7, nt ordered ACGT:0123
);
#
#3D Array for SRp55 scores with nt ordered ACGT:0123
my @SRp55scores = (
[-0.66, 0.39, -1.58, 1.22], #Scores for position 1, nt ordered ACGT:0123
[0.11, -1.58, 0.72, -1.58], #Scores for position 2, nt ordered ACGT:0123
[-0.66, 1.48, -1.58, -0.07], #Scores for position 3, nt ordered ACGT:0123
[0.11, -1.58, 0.72, -1.58], #Scores for position 4, nt ordered ACGT:0123
[-1.58, -1.58, 0.21, 1.02], #Scores for position 5, nt ordered ACGT:0123
[0.61, 0.96, -0.79, -1.58] #Scores for position 6, nt ordered ACGT:0123
);

#************************************************************
#thresholds for SF2, SC35, SRp40, SRp55}
my @thresholds = (1.956, 2.383, 2.760, 2.676);
#************************************************************

#Subroutine to check if inputted sequence is actually DNA sequence, calling on the hash of aa
sub codon2aa{
        my @codon =@_;
        $codonvar = $codon[0];
        $codonvar= uc $codonvar;
        for $key ( keys %genetic_code ) {
                for $i ( 0 .. $#{ $genetic_code{$key} } ) {
                        $cod = $genetic_code{$key}[$i];  #i.e: CGA
                        if ($cod eq $codonvar) {
                        return $key;
                }
                }
        }
  return -1;
```

```
}



#Asking user to input their DNA sequence
print "\t Welcome to The EXONerator.\n\n";
print "\t ******Please input your DNA sequence in the correct reading frame****\n\n";
print "\t Please press 1 to enter your DNA sequnce manually\n";
print "\t Or press 2 to enter your DNA sequence as a file\n";


#Analyzing Entry
$entry=<STDIN>;
chomp($entry); #Takes off the "enter" function from the entered sequence

#Simple if loop for entry of data - input of seq, file or error message
if($entry==1){
	print "Please input your sequence:-\n";
	$dna=<STDIN>;
	chomp($dna);
 } elsif ($entry==2){
	print "Please enter your filename (if file is in another folder, please specify the path): \n";
	$dnafile=<STDIN>;
	open(FILE,$dnafile);
	@filecontents=<FILE>;
	close FILE;
	chomp(@filecontents);
	$dna=join('',@filecontents);
 } else {
	print "Error: You selected an entry other than 1 or 2\n";
	exit;
}
$dna = trim($dna);


#to check if dna is in lengths of 3:
{
	if (length($dna) % 3 != 0) {	#Can use / instead of division sign?
		print "\nYour dna sequence is not a coding sequence - the last codon is less than 3 bases
	long";
		print "\nPlease check your sequence and try again\n";
		print "\nA sequence with line breaks may also cause this error - make sure there are no
	empty lines in your input\n";
		exit;
	}
}
```

```
#Calling on the subroutine to make sure input is only in the form of DNA sequence
{
	my @codon = ($dna =~ m/.../g); #match 3 characters globally across dna sequence

	foreach $codonvar (@codon){
		#print "codon: " . $codonvar . "\n";
		if (codon2aa ($codonvar) == -1) { #looping through the array I just created to check
	inputted dna seq
			print STDERR "Your input included a bad codon: \"$codonvar\"!!\n";
			print "Please check your DNA sequence for errors and try again\n";
			exit;
		}
	}
}


#Subroutine to make sure any changes occurring are only in the correct reading frame. Determines that
	the coding seq is in a particular motif,
# then calls on GETLOWSCORE subroutine to analyze motif
sub changeDna {
	my($motif, $position, $thresh, @scores) = @_;
	print "Motif: " . $motif . "\n";
	$key1="";
	$key2="";
	$startString = "";
	$endString = "";
	if ($position % 3 ==0) {
		$key1 = codon2aa(substr($motif, 0, 3));
		$key2 = codon2aa(substr($motif, 3, 3));
		$endString = substr($motif, 6, length($motif)-6);
		print $startString . "-" . substr($motif, 0, 3) . "-" . substr($motif, 3, 3) . "-". $endString .
	"\n";
		print '----------' . "\n";
	}elsif($position % 3 ==1 ) {
		$startString = substr($motif, 0, 2);
		$key1 = codon2aa(substr($motif, 2, 3));
		if (length($motif) == 8) {
			$key2 = codon2aa(substr($motif, 5, 3));
			print $startString . "-" . substr($motif, 2, 3) . "-" . substr($motif, 5, 3) . "-".
	$endString . "\n";
		}else {
			$endString = substr($motif, 5, length($motif)-5);
			print $startString . "-" . substr($motif, 2, 3) . "-" . $key2 . "-". $endString . "\n";
		}
	}elsif($position % 3 ==2 ) {
		$startString = substr($motif, 0 ,1);
		$key1 = codon2aa(substr($motif, 1, 3));
		if (length($motif) ==6) {
```

```
                $endString = substr($motif, 4, 2);
                print $startString . "-" . substr($motif, 1, 3) . "-" . $key2 . "-". $endString . "\n";
        }elsif (length($motif) ==7) {
                $key2 = codon2aa(substr($motif, 4, 3));
                print $startString . "-" . substr($motif, 1, 3) . "-" . substr($motif, 4, 3) . "-".
    $endString . "\n";
        }elsif (length($motif) ==8) {
                $key2 = codon2aa(substr($motif, 4, 3));
                $endString = substr($motif, 7, 1);
                print $startString . "-" . substr($motif, 1, 3) . "-" . substr($motif, 4, 3) . "-".
    $endString . "\n";
        }
    }
    my ($lowestscoremotif, $lowestscore) = getLowScore($motif, $startString, $key1, $key2,
    $endString, $thresh, @scores);
}


#Subroutine to look for all aa combos with a score below threshold. Randomly chooses one to substitue
    for an above threshold codon and returns it to CHANGEDNA
#If nothing below threshold is found, then it simply returns an empty string.
sub getLowScore() {
    my($motif, $startString, $key1, $key2, $endString, $thresh, @scores) = @_;
    @lowscoresmotif = ();
    @lowscoresscore = ();
    for $i ( 0 .. $#{ $genetic_code{$key1} } ) {
    $cod1 = $genetic_code{$key1}[$i];
        if ($key2 ne "") {
            for $j ( 0 .. $#{ $genetic_code{$key2} } ) {
            $cod2 = $genetic_code{$key2}[$j];
                $newmotif = $startString . $cod1 . $cod2 . $endString;
                $newscore = calculateScore($newmotif, @scores);
                print $newmotif . "" . $newscore . "\n";
                if ($newscore <= $thresh) {
                    print "****Found Low: " . $startString . "-". $cod1 . "-". $cod2 .
    "-" . $endString . "Thresh: " . $thresh . "Score: " . $newscore . "\n";
                    push(@lowscoresmotif, $newmotif);
                    push(@lowscoresscore, $newscore);
                    #return ($newmotif, $newscore);
                }
            }
        }else {
            $newmotif = $startString . $cod1 . $endString;
            $newscore = calculateScore($newmotif, @scores);
            if ($newscore <= $thresh) {
                    print "****Found Low: " . $startString . "-". $cod1 . "-". $cod2 .
    "-" . $endString . "Thresh: " . $thresh . "Score: " . $newscore . "\n";
                    push(@lowscoresmotif, $newmotif);
```

```
                                        push(@lowscoresscore, $newscore);
                                        #return ($newmotif, $newscore);
                                }
                        }
                }
                if (@lowscoresmotif > 0) {
                        print "Random: " . $lowscoresmotif[0] . " " . $lowscoresscore[0] . "\n";
                        my $range = @lowscoresmotif;
                        my $random_number = int(rand($range + 1));
                        print $random_number . " " . $range  ."\n" ;
                        return ($lowscoresmotif[$random_number], $lowscoresscore[$random_number]);
                } else {
                        return ("", "");
                }
}

#Subroutine to calculate the scores of each motif for a given protein
sub calculateScore {
        my($motif, @scores) = @_;
        $score = 0;

        for ($i = 0; $i < length($motif); $i++) {
                if (substr ($motif, $i, 1) eq "A") {
                        $score += $scores[$i][0];
                } elsif (substr ($motif, $i, 1) eq "C") {
                        $score += $scores[$i][1];
                } elsif (substr ($motif, $i, 1) eq "G") {
                        $score += $scores[$i][2];
                } elsif (substr ($motif, $i, 1) eq "T") {
                        $score += $scores[$i][3];
                }
        }
        return $score;
}

#Subroutine to checks all motifs in the block, shifting one base each time. Calls on other subroutines to
        check, calculate and modify.
sub changeDnaScores {
        my($str, $position, $length, $thresh, @scores) = @_;
        $motif = substr ($str,$position,$length);
        if (length($motif) ==$length) {
                $score = calculateScore($motif, @scores);
                #print $motif . " * " . $score . "Thresh: " . $thresh . "\n";
                if($score > $thresh) {
                        my($newmotif, $newscore) =  changeDna($motif, $position, $thresh, @scores) ;
                        if ($newmotif ne "") {
                                $startstr = substr($str, 0, $position);
                                $endlength = length($str) - $position - length($newmotif);
```

```
                                $endstr = substr($str, $position + $length, $endlength);
                #               #print "********* Position: " . $position . "Score" . $score . " Threshold:
        " . $thresh . " new motif: " . $startstr . $newmotif . $endstr . "\n";
                                return $startstr . $newmotif . $endstr;
                        } else {
                                push(@invalidMotif, $motif);
                                push(@invalidMotifPosition, $position);
                                return "";
                        }
                } else {
                        return $str;
                }
        }

}


#Recursive Subroutine, called by SUBPROCESSDNA. Checks each protein score after change, looping
        through each protein again if a change is made
sub ComputeMotifs {
        my($str, $position, $excludeType) = @_;

        $newstr = "";
                print "sf2: \n";
                $sf2newstr= changeDnaScores($str, $position, 7, $thresholds[0], @SF2scores);
                print "sf2: " . $sf2newstr . "\n";
                if ($sf2newstr eq "") {
                        return "";
                } else {
                        $newstr = $sf2newstr;
                }
                print "1" . $newstr . "\n";

                print "SRp40: \n";
                $sr40newstr = changeDnaScores($newstr, $position, 7, $thresholds[2], @SRp40scores);
                print "SRp40: " . $sr40newstr . "\n";
                if ($sr40newstr eq "") {
                        return "";
                }elsif($sr40newstr ne $newstr) {
                        if ($sr40newstr ne $str) {
                                $newstr = ComputeMotifs($sr40newstr, $position, "SRp40");
                                if ($newstr eq "") {
                                        return "";
                                }
                        } else {
                                print "Unable to process this dna properly at position: " . $position .
        "\n";
                                return "";
```

```
                }
        }
        print "2" . $newstr . "\n";

        print "SC35: \n";
        $sc30newstr = changeDnaScores($newstr, $position, 8, $thresholds[1], @SC35scores);
        print "SC35: " . $sc30newstr . "\n";
        if ($sc30newstr eq "") {
                return "";
        }elsif($sc30newstr ne $newstr) {
                if ($sr30newstr ne $str) {
                        $newstr = ComputeMotifs($sc30newstr, $position, "SC35");
                        if ($newstr eq "") {
                                return "";
                        }
                } else {
                        print "Unable to process this dna properly at position: " . $position .
    "\n";
                        return "";
                }
        }
        print "3" . $newstr . "\n";

        print "SRp55: \n";
        $srp55newstr = changeDnaScores($newstr, $position, 6, $thresholds[3],
    @SRp55scores);
        if ($srp55newstr eq "") {
                return "";
        }elsif($srp55newstr ne $newstr) {
                if ($srp55newstr ne $str) {
                        $newstr = ComputeMotifs($srp55newstr, $position, "srp55");
                        if ($newstr eq "") {
                                return "";
                        }
                } else {
                        print "Unable to process this dna properly at position: " . $position .
    "\n";
                        return "";
                }
        }
        print "4" . $newstr . "\n";
    return $newstr;
}

#remove extra white space
sub trim($)
{
    my $string = shift;
```

```
	$string =~ s/^\s+//;
	$string =~ s/\s+$//;
	$string =~ s/\s+//g;
	$string =~ s/(^\s+|\s+$)//g;
	return $string;
}

#Calls on the recursive subroutine; processes the 16 block fragment and returns modified block to
	SUBPROCESSDNA
sub processSubdna {
	my($str) = @_;
	$previousSubdna = $str;

	$count =0;
	while(($count == 0) || ($str ne $previousSubdna && count < 11)) {
		@invalidMotif = ();
		@invalidMotifPosition = ();
		print "**********************************************************" . "\n";
		print $str . "\n";
		$previousSubdna = $str;

		$position = 0;
		while ($position < length($str)) {
			print "Position: " . $position . " str: " . $str. "\n";
			$new = ComputeMotifs($str, $position) . "\n";
			$new = trim($new);
			if (($new ne "") && ($new ne $str)) {
				$str = $new;
			}
			$position++;
			print "----------------------------" . "\n";
		}
		$count++;
	}
	return $str;
}

#Processed the entire DNA fragment; after making a copy of the original DNA to modify, replaces
	modified blocks in cop from PROCESSUBDNA and shifts 6 positions
#each time to make the new block, which is sends to PROCESSUBDNA.
sub processDna {
	my ($dna, $globalPosition) = @_;
	$next = "true";
	$newdna = $dna;
	while ($next eq "true") {
		print "Global Position:" . $globalPosition . "\n";

		$restDna = substr($newdna, $globalPosition);
```

```
                print  "Rest Dna: " . $restDna . "Length: " . length($restDna) . "\n";
                if (length($restDna) <16) {
                        $subdna = substr($newdna,  $globalPosition);
                        $next = "false";
                } else {
                        $subdna = substr($newdna,  $globalPosition, 16);
                }

                $modifiedSubdna = processSubdna($subdna);
                $startDna = substr($newdna, 0, $globalPosition);
                $seposition = $globalPosition + length($modifiedSubdna);
                print $seposition . "\n";
                $endDna = substr($newdna, $seposition);

                print $startDna . "-" . $modifiedSubdna . "-" . $endDna . "\n";
                $newdna = $startDna . $modifiedSubdna . $endDna;

                $globalPosition = $globalPosition + 6;
                print "Next: " . $next . " " . $globalPosition;
        }
        return $newdna
}
$dna = uc $dna;
$dna = trim($dna);
@invalidMotif = ();
@invalidMotifPosition = ();

$modifiedDna = processDna($dna, 0);

print $dna . "\n";
print "**********************************************************" . "\n\n";
print "Inputted Dna Sequence       : " . $dna . "\n";
if ($modifiedDna ne $dna) {
    print "DNA has changed. DNA with silenced ESEs: " . $modifiedDna . "\n";
} else {
    print "DNA has not changed" . "\n";
}
print "\n";

if (@invalidMotif> 0) {
        print "List of Motifs with scores higher than the threshold that could not be changed: \n";
        for ($i=0; $i<@invalidMotif; $i++) {
                print $invalidMotif[$i] . " Position: " . $invalidMotifPosition[$i] . "\n";
        }
}
```