GENOME-WIDE APPROACHES FOR IDENTIFYING NOVEL GENETIC AND EPIGENETIC EVENTS IN GASTROINTESTINAL CANCER

by

RYAN EVERETT FECTEAU

Submitted in partial fulfillment of the requirements
For the degree of Doctor of Philosophy

Dissertation Advisor: Sanford Markowitz, M.D., Ph.D.

Department of Pathology
CASE WESTERN RESERVE UNIVERSITY

August 2015
CASE WESTERN RESERVE UNIVERSITY
SCHOOL OF GRADUATE STUDIES

We hereby approve the thesis/dissertation of

Ryan Everett Fecteau

candidate for the Doctor of Philosophy degree *.

(signed) Mark Jackson (chair of the Committee)
Sanford D. Markowitz
Peter Scacheri
Zhenghe Wang
William Schiemann
Clive R. Hamlin

June 19th, 2015

*We also certify that written approval has been obtained for any proprietary material contained therein.
Dedication

To my parents, David and Rita Fecteau, whose love and support has motivated me to pursue my ambitions and suffused my character with their spirit of generosity and kindness.
TABLE OF CONTENTS

LIST OF TABLES iv

LIST OF FIGURES v

ACKNOWLEDGEMENTS vii

ABSTRACT 1

Chapter 1: Targeted Pyrosequencing of GNAS Mutations in Colorectal Cancer Tumors

Identifies a Unique Molecular and Histologic Phenotype 3

1.1 Introduction 3

1.1.1 Epidemiology of Colorectal Cancer 3

1.1.2 Genetic Etiology of Colorectal Cancer 3

1.1.3 The Genomic Landscape of Somatic Mutations in Colorectal Cancer 6

1.1.4 Mining the Hills Identifies Recurrent Mutations in the Oncogene GNAS 8

1.1.5 The GNAS Gene and Canonical GNAS Function 9

1.1.6 GNAS in Cancer 11

1.2 Approach 12

1.3 Abstract 13

1.4 Materials and Methods 14

1.5 Results 17

1.5.1 Frequency of GNAS Hotspot Mutations in Colorectal Cancer 17

1.5.2 GNAS mutations are associated with a villous morphology 17

1.5.3 GNAS mutant CRCs are predominantly KRAS/BRAF mutant and located in the proximal colon 18

1.6 Conclusions 19
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.7 Discussion and Future Directions</td>
<td>20</td>
</tr>
<tr>
<td>1.8 Figures and Tables</td>
<td>26</td>
</tr>
<tr>
<td><strong>Chapter 2. Genetic Study of Familial Barrett's Esophagus and Esophageal</strong></td>
<td></td>
</tr>
<tr>
<td>Adenocarcinoma Identifies <em>VSIG10L</em> as a Candidate Disease Susceptibility Gene</td>
<td>33</td>
</tr>
<tr>
<td>2.1 Introduction</td>
<td>33</td>
</tr>
<tr>
<td>2.1.1 Study of Familial Cancer Syndromes</td>
<td>33</td>
</tr>
<tr>
<td>2.1.2 Next Generation Sequencing Overview</td>
<td>34</td>
</tr>
<tr>
<td>2.1.3 Whole Exome Sequencing for Mutation Discovery in Familial Cancer</td>
<td>35</td>
</tr>
<tr>
<td>2.1.4 Familial Barrett's Esophagus</td>
<td>38</td>
</tr>
<tr>
<td>2.2 Approach</td>
<td>40</td>
</tr>
<tr>
<td>2.3 Abstract</td>
<td>40</td>
</tr>
<tr>
<td>2.4 Materials and Methods</td>
<td>41</td>
</tr>
<tr>
<td>2.5 Results</td>
<td>49</td>
</tr>
<tr>
<td>2.6 Conclusions</td>
<td>52</td>
</tr>
<tr>
<td>2.7 Discussion and Future Directions</td>
<td>53</td>
</tr>
<tr>
<td>2.8 Figures and Tables</td>
<td>59</td>
</tr>
<tr>
<td><strong>Chapter 3: Interrogating the Colon Cancer Methyome for Biomarker Discovery</strong></td>
<td>72</td>
</tr>
<tr>
<td>3.1 Introduction</td>
<td>72</td>
</tr>
<tr>
<td>3.1.1 Epigenetics Overview</td>
<td>72</td>
</tr>
<tr>
<td>3.1.2 DNA Methylation</td>
<td>72</td>
</tr>
<tr>
<td>3.1.3 DNA methylation in colorectal cancer</td>
<td>74</td>
</tr>
<tr>
<td>3.1.4 Methylated tumor DNA as a Biomarker</td>
<td>75</td>
</tr>
<tr>
<td>3.1.5 Reduced Representation Bisulfite Sequencing: a tool for examining genome-wide DNA methylation</td>
<td>77</td>
</tr>
<tr>
<td>3.2 Approach</td>
<td>79</td>
</tr>
</tbody>
</table>
3.3 Materials and Methods 79

3.4 Results 86

3.4.1 Reduced Representation Bisulfite Sequencing and Variant Filtering to Identify Candidate Methylated Loci 86

3.4.2 Patch Validation identifies 4 primary methylated candidate biomarkers 87

3.4.3 Window Identification in top candidate biomarkers 88

3.4.4 Validation of Candidate Windows in an independent tumor cohort 88

3.5 Conclusions 89

3.6 Discussion and Future Directions 89

3.7 Figures and Tables 93

Appendix 103

References 110
LIST OF TABLES

Table 1.1 Tumor molecular and clinical characteristics according to GNAS mutation status ..............................................................................................................28

Table 1.2 Clinical and molecular characteristics of GNAS mutant tumors ..........................................................................................................................30

Table 1.3 GNAS mutation frequency in adenomas .................................................................................................................................32

Table 1.4 Characteristics of GNAS mutant adenomas ...........................................................................................................................32

Table 2.1 Clinical Characteristics of members from FBE family 31 tested for VSIG10L S631G germ-line variant ..............................................................................................65

Appendix Table 1 Primer sequences and coordinates for patch amplification for targeted resequencing ............................................................................101

Appendix Table 2 Window primer sequences ..................................................................................................................106
LIST OF FIGURES

Figure 1.1 Genetic pathways frequently targeted in colon cancer ......................26
Figure 1.2 Somatic mutation landscape of CRC ..............................................27
Figure 1.3 GNAS codon 201 mutations.................................................................29
Figure 1.4 GNAS tumors are associated with a villous morphology .................31
Figure 2.1 Schematic of next generation sequencing ...........................................59
Figure 2.2 Workflow schematic of whole exome library preparation ..................60
Figure 2.3 Approaches for familial sequencing studies .......................................61
Figure 2.4 Illustration of anatomic and histologic features of Barrett’s esophagus ..................................................................................................................62
Figure 2.5 VSIG10L segregation, multiple sequence alignment, and expression in BE associated lesions ..................................................................................64
Figure 2.6 Schematic of variant filtering workflow ..............................................67
Figure 2.7 Representative chromatograms of Sanger sequencing .......................68
Figure 2.8 Photomicrographs and immunofluorescence of organotypic cell culture model expressing WT and S631G variant ............................................70
Figure 2.9 Photomicrographs of organotypic cell culture model expressing VSIG10L G769S somatic variant ........................................................................71
Figure 3.1 Overview of RRBS library preparation ...............................................93
Figure 3.2 Variant filtering workflow for candidate CpGs .................................94
Figure 3.3 Diagram illustrating patches, PCR amplicons, and windows ..........95
Figure 3.4 Schematic of methylation per-read versus methylation per CpG analysis .................................................................96

Figure 3.5 Sensitivity and specificity graphs of top 4 candidate patches in targeted resequencing expanded discovery set .................................................................97

Figure 3.6 Sensitivity and specificity graphs of top 4 candidate windows from targeted resequencing in expanded discovery set.................................98

Figure 3.7 Sensitivity and specificity graphs of top 4 window performance in validation set.........................................................................................99
ACKNOWLEDGEMENTS

I would like to begin by thanking those that helped me arrive at this stage of my training. Starting with my high school science teacher Mr. Art Hawley, my undergraduate advisor Dr. William (Bill) Marcotte, and my professors at Clemson University that offered guidance along the way. Of course I have to give special thanks to my research mentor at Clemson, Dr. James Morris, who not only oversaw my research efforts, but also taught me some of life’s harder lessons. Jim, I’ll always take the high road.

I will be forever grateful for the support I received from the CWRU MSTP and for being given the opportunity to train at this institution. It goes without saying the gratitude I will always have for my advisor, Dr. Sanford Markowitz, for taking a chance on me during a trying period of personal recovery. Not only did he provide instruction and encouragement, but also reinvigorated my confidence that had been sorely absent at this time. He has transformed my scientific thinking and the manner in which I approach provoking questions, undoubtedly improving my capabilities as a researcher. The magnitude of indebtedness I feel is beyond description.

Lastly, few goals are accomplished without the love and support of family and friends. My parents, to whom this work is dedicated, have been my greatest advocates since my birth. They have pushed for me to focus on my academics and have provided financial and moral support all along the way. I can only hope to begin to repay them one day. Thank you to my sister Megan who began as my tutor at a young age and is now one of my closest friends helping me navigate this life. Also to Annie, whose recent emergence into my life has been nothing short of profound. Your extraordinary patience and composure was so crucial during this time. I am incredibly thankful to have you.
Cancer is a disease caused by the sequential acquisition of genetic and epigenetic changes that transform a normal cell into an ungovernable malignancy. Advances in technologies to interrogate these changes on a genome-wide scale have reshaped our understanding of this disease and rapidly expanded the field of cancer genomics. The objective of the projects described herein was to adapt high-throughput methods to discover and characterize novel genetic and epigenetic events in gastrointestinal cancers with the intention of translational application.

Three primary studies comprise the body of this thesis. In the first, we identify recurrent mutations in the oncogene GNAS found by sequencing of the colorectal cancer genome. Targeted pyrosequencing of GNAS in a colon cancer tumor cohort established a mutation frequency in colorectal cancer of around 2%. By comparing GNAS mutation status with clinical and pathologic data, we found that GNAS mutations associate with a distinct class of colon cancers that are characterized by concomitant KRAS or BRAF mutations, by location in the proximal colon, and by a unique association with a villous morphology.

The second study applied a targeted next generation sequencing approach, whole exome sequencing, to identify a germ-line susceptibility variant in a familial syndrome of Barrett’s esophagus and esophageal adenocarcinoma. Sequencing revealed a private
variant in the uncharacterized gene VSIG10L that segregated with disease in a large familial Barrett’s esophagus family and functional studies suggested the discovered variant disrupted maturation of the normal esophageal epithelium. Lastly, we describe using reduced representation bisulfite sequencing to examine the colon cancer methylome for candidate methylated biomarker loci. Preliminary data from this work has identified 3 candidate methylated DNA biomarkers for potential use in detection of colorectal cancer in patient plasma.

Cumulatively, these studies have identified a rare mutation in colon cancer associated with a unique molecular phenotype, a private germ-line variant predisposing to familial Barrett’s esophagus and adenocarcinoma, and candidate methylated biomarkers for detection and monitoring of colon cancer, all of which are of potential clinical importance.
CHAPTER 1: TARGETED PYROSEQUENCING OF GNAS MUTATIONS IN COLORECTAL CANCER TUMORS IDENTIFIES A UNIQUE MOLECULAR AND HISTOLOGIC PHENOTYPE

1.1 Introduction

1.1.1 Epidemiology of Colorectal Cancer

Colorectal cancer (CRC) is the second leading cause of cancer related death in United States and it is estimated that over 50,000 U.S. men and women will die from this disease in 2015[1]. Though incidence is on the decline, CRC remains 3rd overall in annual incidence for men and women combined with over 130,000 cases in 2015. Improvements in screening and treatment have raised the 5-year survival rate to 65%, however 20% of newly diagnosed cases will present with late-stage, metastatic disease that carries a dismal 5-year survival rate of 11%.

1.1.2 Genetic Etiology of Colorectal Cancer

Sporadic colon cancer, as in many other cancers, occurs as a result of genetic alterations that include inactivation of tumor suppressor genes, constitutive activation of oncogenes, and deregulation of gene expression. The progression of CRC from an adenoma to an invasive tumor is often accompanied by a sequence of gene mutations occurring at specific stages of neoplasia (Figure 1.1). This step-wise transformation, often referred to colloquially as the “Vogelgram” in reference to Bert Vogelstein who conceived of the paradigm[2], has since been extended to model numerous other solid tumors and illustrates the genetic basis underlying cancer development [3–5].
Broadly, CRC can be divided into two distinct subtypes based on the status of the DNA mismatch repair (MMR) pathway. Cancers with deficient MMR are grouped as microsatellite instable (MSI) as they exhibit a higher rate of mutation and expansion of microsatellite repeats throughout the genome. These MSI cancers can be hereditary due to either germ-line mutation in mismatch repair genes or sporadic due to siliencing by epigenetic DNA methylation. Microsatellite stable (MSS) cancers usually arise in the context of genomic instability due copy number changes and structural aberrations in chromosomes. Chromosomal instability is an important mechanism for loss of heterozygosity (LOH) of tumor suppressor genes in these cancers.

The initiating event in colon cancer is largely considered to be activation of the Wnt signaling pathway[6]. Canonical Wnt signaling commences upon secreted Wnt ligands binding their cognate receptor, frizzled, leading to β-catenin accumulation in the cytoplasm. This then allows for β-catenin translocation to the nucleus, where it may associate with nuclear proteins from the T-cell factor and lymphocyte enhancer factor (TCF/LEF) families to form a transcription factor capable of modulating the expression of genes important in cell proliferation, migration, and differentiation[7]. The most common genetic aberrations leading to activation of this pathway are truncating mutations of the tumor suppressor gene APC. Germ-line mutations in APC were discovered as the cause of familial adenomatous polyposis (FAP), a rare inherited syndrome characterized by the development of hundreds of colonic polyps and the inevitable progression to adenocarcinoma. The function of the APC gene is to sequester β-catenin in a complex with the proteins axin and GSK3β in the cytoplasm and facilitate its ubiquitin-mediated degradation. Mutations in APC that abrogate this function lead to
an accumulation of β-catenin in the cytoplasm and increased concentrations in the nucleus where it can exert its proliferative effects. Biallelic inactivation of APC by mutation or loss of heterozygosity (LOH) occurs early on in colon cancer progression and has been detected in about 50% of aberrant crypt foci and 80% of small adenomas (<1cm³)[7,8].

Progression of a small adenoma to a large adenoma (>1cm³) is associated with mutations that activate the oncogenes KRAS and BRAF, with respective mutational frequencies of approximately 40% and 10% in CRC[9]. These mutations lead to constitutive signaling through the RAS-RAF-MAPK pathway that induce cell proliferation via upregulation of transcription factors required for cell cycle entry and progression[10,11]. Oncogenic KRAS also activates the PI3K-AKT pathway that can regulate proteins involved in cell motility and apoptosis in addition to proliferative effects[12]. Mutations activating PI3K are also common in colon cancer and have been shown to promote growth and invasion both in vitro and in vivo[6,13]. The transition from adenoma to carcinoma is often accompanied by inactivating mutations in two tumor-suppressor genes: TP53 and transforming growth factor beta receptor II (TGFBRII)[14–16]. The TP53 gene resides on chromosome 17p and encodes for the tumor-suppressor p53 that acts to initiate cell cycle arrest and apoptosis in response to a number of exogenous and endogenous stresses[17]. In CRC, the TP53 gene usually harbors one inactivating mutation while the other is usually lost as a result of chromosome 17p deletion while LOH or mutation is infrequent in earlier lesions[18]. Similarly, mutation of the TGFBRII gene is most often detected in CRC with microsatellite instability (MSI) and not in earlier MSI positive adenomas[15]. The TGF-β
receptor II is a potent inhibitor of colonic epithelial cell growth and mediates its effects through phosphorylation and activation of SMAD transcription factors[19]. Though mutations in TGFBRII in microsatellite stable (MSS) CRC occur at a lower frequency, approximately 50% of MSS tumors have abolished TGF-β signaling due to mutations and/or deletions in the downstream SMAD transcription factors [16,20].

1.1.3 The Genomic Landscape of Somatic Mutations in Colorectal Cancer

As recent as the early 2000’s, identification of genes involved in disease pathogenesis was a laborious and tedious process, if not fortuitous. The genes in pathways discussed above were discovered using a combination of approaches from classical genetic cancer research including cytogenetic analysis and positional cloning studies of familial and sporadic cancer, comparison of human genes with transforming viral homologues, and biological assays that either demonstrated transforming capabilities of putative oncogenes or growth suppressive effects of tumor suppressor genes[21,22]. The sequencing and annotation of the human genome revolutionized the genetic study of human disease as it provided a map of the normal DNA code as a reference for comparison of the genetic changes that predispose or accompany a disease state. In no other disease is this more evident than in cancer, given its genetic basis of acquired, somatic mutation.

Preceding the emergence and adaptation of next-generation sequencing technology, a pair of landmark studies published in 2006 and 2007 sequenced the full set of protein-coding DNA in colon cancer in an effort to identify other commonly mutated genes and define the complement of gene mutations that constitute CRC[23,24]. From these studies three pivotal conclusions were drawn. First, the set of genes that were found
to be recurrently mutated and thus most likely to be driver mutations in CRC comprised 140 cancer candidate (CAN) genes. Secondly, while CRC tumors harbored on average around 90 non-silent mutations per sample, only about 15 (17%) of these mutations occurred in CAN genes, suggesting the remaining were likely passenger mutation events not involved in carcinogenesis. Lastly, mutations in CAN genes were of two varieties: those occurring at high frequency in CRC, referred to as “mountains” on the genomic landscape, and those occurring at a lower frequency (<5%) referred to as “hills” (Figure 1.2).

The mountains of the landscape were scarce, as only a few genes were mutated with high frequency in all tumors. These included most of the well-defined tumor suppressor and oncogenes characterized in CRC thus far such as APC, p53, and KRAS. The hills that dominated the landscape are perhaps the defining conclusion of this study. The finding that most of the mutations of the 140 CAN genes in CRC occur at low frequency illustrated the remarkable heterogeneity of the mutational landscape in CRC and implicated numerous avenues for disrupting cellular growth that could converge onto cancer. Indeed, since this study the rapid adaptation of next generation sequencing has allowed for genome sequencing of countless other cancers that has yielded similar findings: a mutation landscape with few mountains and dominated by hills[25].

While many of the low frequency mutations occur in genes of known function or affect genes in common cancer-associated pathways, the contribution of a given mutation to tumorigenesis can vary by cancer type. For example, NOTCH signaling is an important pathway in differentiation and tissue homeostasis mutated frequently in certain leukemias, non-small cell lung carcinoma (NSCLC), and head and neck squamous cell
carcinoma (HNSCC)[26–29]. However, in NSCLC and leukemia NOTCH is acting as an oncogene and these mutations are activating whereas in HNSCC these mutations appear to be inactivating and indicate NOTCH is acting as a tumor suppressor in this cancer type. Observations such as these illustrate the importance of further investigating the cancer specific function of a given gene mutation.

1.1.4 Mining the Hills Identifies Recurrent Mutations in the Oncogene GNAS

In addition to functional characterization, low frequency mutations can be studied to determine mutation effects on clinical outcome and cancer phenotype. Cancer subtypes with distinct molecular and histologic phenotypes can be associated with unique somatic mutation profiles. These associations have often been observed with single gene mutations being linked to a given histologic or molecular phenotype, such as CDH1 inactivating mutations in diffuse gastric adenocarcinoma, activating BRAF mutations in sessile serrated adenomas and right-sided CRC, and polymerase epsilon (POLE) mutations in hypermutated CRC[30,31]. Such molecular-phenotypic relationships are useful for classifying and studying disease course in the clinic and determining possible associations with clinical outcomes.

Bearing this in mind, we surveyed the 140 CAN genes identified in CRC for candidate genes to conduct a study examining somatic mutation frequencies in a large CRC tumor cohort with the aim of correlating mutations with clinical and histopathologic characteristics. From this survey we honed in on an interesting candidate oncogene, GNAS, for numerous reasons. At the time, GNAS had been recently found to be frequently mutated in intraductal papillary mucinous neoplasms (IPMN) of the pancreas,
a malignant precursor to pancreatic adenocarcinoma[32]. Notably, these mutations often co-occurred with KRAS activating mutations, which is commonly mutated in CRC[6]. Preliminary data also showed that of the GNAS mutations observed, all were in late-stage metastatic disease, raising the possibility that this mutation either occurred in a later stage or conferred a more aggressive phenotype. Lastly, due to GNAS being an oncogene, it is inherently a putative candidate for targeted therapy. Given the description of GNAS mutations in other GI neoplasms, its association with the commonly mutated CRC gene KRAS, and a possible association with metastatic disease, we elected GNAS for further scrutiny of its somatic mutation frequency in CRC and the corresponding pathology of GNAS mutant tumors.

1.1.5 The GNAS Gene and Canonical GNAS Function

The GNAS gene lies in a complex, imprinted locus on chromosome 20 harboring multiple promoters and giving rise to multiple gene products. The predominant gene product is the Gαs subunit that tethers G-protein coupled receptors (GPCRs) to adenylate cyclase[33]. Transcription of Gαs begins from the most downstream promoter in the locus and the transcript is biallelically expressed in all tissues. The other products expressed off this locus include the oppositely imprinted genes NESP55 and XLαs, which use alternate promoters approximately 45 and 30 kb upstream of the Gαs exon1, respectively[34]. The NESP55 gene product is structurally and functionally unrelated to Gαs, utilizing its alternate first exon as the sole coding region and is expressed off the maternal allele. XLαs is a neuroendocrine-specific Gαs isoform, also differing in composition by an alternate first exon and expressed off the paternal allele. Multiple endocrine disorders and
clinical syndromes result from genetic aberrations at this locus. Activating mutations in $G_\alpha_s$ have been found in many tumors and as the cause of McCune-Albright syndrome; inactivating mutations cause Albright hereditary osteodystrophy; and imprinting defects result in pseudohypoparathyroidism type b[33].

The major product of the GNAS locus is the $G_\alpha_s$ subunit of heterotrimeric G-proteins that acts to transduce signals from GPCRs to the effector enzyme adenylate cyclase in the G-stimulatory (Gs) pathway, leading to the production of cAMP. This 45-kD protein consists of a GTPase domain that includes sites for GTP binding and receptor/effector interactions and a helical domain involved in maintaining bound GTP. In the resting state, $G_\alpha_s$ is bound with GDP and complexed with $\beta\gamma$-subunits. Ligand-bound GPCRs interact with $G_\alpha_s$ and promote the exchange of bound GDP with free GTP, inducing a conformational change that allows the $G_\alpha_s$ to dissociate from the $\beta\gamma$-subunits and activate adenylate cyclase through direct interaction[35,36]. The principal mediator of cAMP signaling is the ubiquitous protein kinase A (PKA), a heterotetramer composed of two catalytic subunits and two regulatory subunits. Upon cAMP binding to the regulatory subunits, they undergo a conformational change and release the catalytic subunits, allowing them to phosphorylate cytoplasmic targets and the transcription factor cAMP-response element-binding protein (CREB)[37]. These events have many cellular consequences and regulate a multitude of pathways including glucose metabolism in hepatocytes and skeletal myocytes, lipid metabolism in adipocytes, water and ion transport in renal cells, and contraction in smooth myocytes. The effects of cAMP on cell proliferation are varied depending on the cell type and observations of mitogenic and anti-proliferative effects have been noted in many systems[38,39]. In addition, the $\beta\gamma$-
subunits have been shown to interact with and activate the PI3K signaling pathway, implicating another pathway commonly exploited by proliferating tumor cells[40].

1.1.6 GNAS in Cancer
Activating mutations in GNAS were first discovered in growth hormone secreting pituitary tumors and have since been reported in a number of other endocrine tumors, Leydig cell tumors, intraductal papillary mucinous neoplasms (IPMN) of the pancreas, and cholangiocarcinoma [41–46]. Almost all observed activating mutations are missense mutations occurring at two highly conserved residues: Arg\textsuperscript{201} and Gln\textsuperscript{227}. Both of these residues lie in the GTPase binding domain and are important for the intrinsic GTPase activity of Go\textsubscript{s}. Most commonly are substitutions at Arg\textsuperscript{201} with either cysteine (R201C) or histidine (R201H) and both have been shown to inhibit the GTPase activity of Go\textsubscript{s} and cause constitutive activation[42]. In pituitary and other endocrine tumors the link between cAMP synthesis and cell growth is apparent as growth and hormone release are regulated by trophic hormones that activate Go\textsubscript{s}-cAMP pathway[47]. In other tissues however, the connection is not as easily discernable since cAMP can have anti-proliferative or proliferative effects depending on the cell type[38].

Insights into how GNAS might be oncogenic were gleaned from the role of prostaglandins in CRC. Prostaglandins are signaling molecules synthesized from arachadonic acid that have multiple physiologic roles. The prostaglandin E2 (PGE\textsubscript{2}) is a pro-inflammatory metabolite whose receptors, EP2 and EP4, are GPCRs that signal through the Go\textsubscript{s}-cAMP pathway. Nonsteroidal anti-inflammatory drugs (NSAIDs), which inhibit the enzymes cyclooxygenase-1 (COX-1) and cyclooxygenase-2 (COX-2) involved
in prostaglandin biosynthesis, can reduce the number and size of polyps in patients with FAP[48,49]. These anti-tumorigenic effects of NSAIDs in CRC have been attributed to decreased PGE$_2$ concentrations in the colonic mucosa, as PGE$_2$ has been shown to be a potent mitogen in colon cancer cells[50]. The mechanism whereby PGE$_2$ activation of the G$\alpha_5$-cAMP pathway induces cell proliferation has been proposed to be through activation of the $\beta$-catenin axis by disrupting the formation of the $\beta$-catenin cytoplasmic degradation complex. This occurs by direct interaction of G$\alpha$ with axin, blocking GSK3$\beta$ from phorsphorylating $\beta$-catenin[51]. Concomitant activation of the PI3K-Akt pathway by the free $\beta\gamma$-subunits leads to phosphorylation and inactivation of GSK3$\beta$. In mouse models of colon cancer, introduction of a constitutively active GNAS allele into APC$^{\text{min/+}}$ mice was shown to further promote intestinal tumorigenesis and augment signaling through the Wnt/$\beta$-catenin and MAPK pathways[52]. While much evidence has accumulated illustrating the oncogenic function of GNAS activating mutations in CRC, no study has examined a potential role for GNAS in metastasis. Interestingly, upregulated COX-2 has been shown to enhance liver metastasis in CRC mouse models and this phenotype can be reversed by pretreatment with COX inhibitors [53,54], indicating increased activation of G$\alpha_5$ by elevated levels of PGE$_2$ might confer a more aggressive phenotype.

1.2 Approach

Reported frequencies of GNAS activating mutations in CRC have differed among various groups, ranging from as little as 0.5% to 9% [52,55–57]. In the present study, we employed a sensitive pyrosequencing platform to sequence the codon R201 mutational hotspot in a cohort of 428 sporadic colon tumors to ascertain a more precise frequency in
CRC. We also assayed for mutations in KRAS and BRAF to determine the prevalence of coincident GNAS/KRAS or GNAS/BRAF mutations in our tumor cohort. Clinical and pathologic data was reviewed to determine if GNAS mutant tumors are associated with any unique clinical or morphologic phenotypes.

1.3 Abstract

The purpose of this study is to determine the genetic frequency of GNAS activating mutations in colorectal cancer and the corresponding pathology of GNAS mutant tumors. Oncogenic mutations in GNAS have been described in a number of neoplasms including those of the pituitary, kidney, pancreas, and, more recently, in colon cancer. To ascertain the frequency in colon cancer we employed a sensitive pyrosequencing platform for mutation detection of the R201C and R201H GNAS hotspots in tumor samples representing all clinical stages. We additionally assayed for KRAS and BRAF mutations as previous reports have shown that these often co-occur with activating GNAS mutations. Of the 428 colon tumors assayed, mutations in GNAS were present in 10 of the samples (1.3%), indicating this is a significant, albeit infrequent, mutation in colorectal tumors. Nine GNAS mutant tumors (90%) harbored concomitant activating mutations in either the KRAS or BRAF oncogene, which was significantly greater than the mutation frequency of these genes in the tumor population (56%, p <0.0305). All ten of the GNAS mutant tumors arose in the right (proximal) colon (p < 0.007), and 7 of 8 reviewed cases exhibited a marked villous morphology. Taken together, these data indicate that GNAS mutant colon tumors commonly have synchronous mutations in
KRAS or BRAF, are right-sided in location, and are associated with a villous morphology.

1.4 Materials and Methods

Ethics Statement

The tumor sample accrual protocol entitled, “CWRU 7296: Colon Epithelial Tissue Bank”, was approved by the University Hospitals Case Medical Center Institutional Review Board for Human Investigation with the assigned UH IRB number 03-94-105. Under this protocol, discarded tissue was obtained through written informed consent from patients for research use.

Tumor Specimens

Tumor specimens were obtained from a frozen archive that consisted of 428 unselected colorectal cancers without reported family history (hereafter referred to as sporadic colorectal adenocarcinomas) accrued under the above protocol. Clinical data was obtained and assembled through individual pathology case reports for each tumor. Microscopic review of tumor morphology for selected samples was performed by an anatomic pathologist (J.W). All samples were assayed for mutations in GNAS codon 201, KRAS codons 12, 13, 61, and 146, and BRAF codon 600 using both Sanger sequencing and pyrosequencing.

DNA extraction and MSI testing

14
Genomic DNA extraction from tumor samples was performed using either a standard guanidine thiocyanate protocol[58] or the DNeasy Blood and Tissue Kit (QIAGEN). Testing for MSI status at genomic loci BAT26 and BAT40 was performed as previously described[59].

Pyrosequencing

Pyrosequencing analysis of GNAS codon 201 was performed on all samples. Pyrosequencing assays were designed using the PSQ Assay Design software (QIAGEN, Chatsworth, CA) that included GNAS codon 201, KRAS codons 12 and 13, KRAS codon 61, KRAS codon 146, and BRAF codon 600. For each assay, one of the PCR primers was biotinylated at the 5’ end and purified using high performance liquid chromatography.

Primer sequences are as follows. GNAS: For: 5’-biotin- TTGGCTTTGGTGAGATCCATTG-3’, Rev 5’-CACCTGGAACCTTGGTCTCAAGAT-3’, Seq 5’- TTCCAGAAAGTCAGGACA-3’;

KRAS codons 12 and 13: For 5’- TCGATGGAGGAGTTTGTAAATGA-3’, Rev 5’- biotin-TTCGTCCACAAAATGATTCTGA-3’, Seq 5’-CTTGTTGGTAGTTGGGAGC-3’;

KRAS codon 61: For 5’- CAGACTGTGTTTCTCCCTTCTCA-3’, Rev 5’- biotin-TCCTCATGTACTGGTCCCTCATTG-3’, Seq 5’- ATATTCTCGACACAGCAG-3’;

KRAS codon 146: For 5’- AGGCTCAGGACCTTAGCAAGAGTT-3’, Rev 5’-biotin-GCCCTCTCAAGAGACAAAAACAT-3’, Seq 5’-AATTCCTTTTTATTGAACAT-3’;

BRAF codon 600: For 5’- TTCATGAAGACCTCACAGTAAAAA-3’, Rev 5’-biotin-CCACAAAATGGATCCAGACA-3’, Seq 5’- TGATTTTGGTCTAGCTACA-3’. All PCR reactions were performed using FastStart Taq (Roche) and primer concentrations of 0.2 uM. Cycling conditions included an initial denaturation step at 95 C for 4 min, and 49
cycles of 95°C for 15 s, 54°C for 30 s, and 72°C for 20 s. Following PCR, amplification products were sequenced on a PyroMark MD pyrosequencing instrument (QIAGEN) and mutation analysis was conducted as previously described[60].

**Sanger Sequencing**

Sanger sequencing was used to confirm all mutations detected by pyrosequencing analysis. Isolated genomic DNA from tumor samples was used for PCR amplification of regions encompassing codon 201 of GNAS, codons 12, 13, 61, and 146 of KRAS, and codon 600 of BRAF. Forward and reverse primers used for PCR amplification were tagged with a 5’ M13 forward (5’-GTAAACGACGGCCAGT-3’) and 5’ M13 reverse (5’-CAGGAAACAGCTATGAC-3’) universal primer sequence, respectively. Primer sequences were as follows. GNAS: For 5’- GTTGGCAATTGTGAGC-3’, Rev 5’- CCCTGATCCAACACACAG-3’; KRAS codons 12 and 13: For 5’- TGTTGGAGTATTTGATAGTGTA-3’, Rev 5’- CATGAAAATGGTCAGAGAA-3’; KRAS codon 61: For 5’- TCCAGACTGTGTTTCTCCCT-3’, Rev 5’- AACCCACCTATAATGGTGAATATCT-3’; KRAS codon 146: For 5’- AGAAGCAATGCCCTCTCAAAG-3’, Rev 5’-GGACTCTGAAGATGTACCTATGGTC-3’; BRAF codon 600: For 5’- TCATAATGCTTGCTCTGATAGGA-3’, Rev 5’- GGCCAAAATTTAATCAGTGGA-3’. All reactions were carried out using 0.4 μM concentration of each primer and FastStart Taq polymerase (Roche, Indianapolis, IN). Cycling conditions for all primer pairs consisted of an initial denaturation at 95°C for 4 min followed by 39 cycles of 95°C for 30 s, 58°C for 30 s, 72°C for 30 s, and a final elongation at 72°C for 3 min.

**Statistical Analysis**
Fisher’s exact test was used to assess differences in the proportion of GNAS mutant tumors between classes of gender, ethnicity, KRAS/BRAF status, microsatellite stability status, clinical stage, and tumor location. A two-tailed P-value of less than 0.05 was considered significant.

1.5 Results

1.5.1 Frequency of GNAS Hotspot Mutations in Colorectal Cancer
Pyrosequencing detected activating mutations of GNAS codon 201 in ten of the 428 (1.3%) colorectal adenocarcinomas (Table 1.1). Each of these mutations was also validated using Sanger sequencing (Fig 1.3). Of the ten GNAS codon 201 mutations detected, we identified seven p.R201H and three p.R201C amino acid substitutions, all of which were mutually exclusive (Table 1.2). Seven of these mutations detected arose among the 377 microsatellite stable tumors tested (1.9%), and three arose among the 41 tumors with microsatellite instability (7.7%). The increased frequency of GNAS mutations in microsatellite unstable tumors was of borderline statistical significance (P=0.065). The mutation frequency at codon 201 of 0.0063 per diploid base pair is significantly higher than the background mutation rate of 1.2x10^{-6} mutations per base pair that typifies microsatellite stable colon cancers (P=3E(-36))[16].

1.5.2 GNAS mutations are associated with a villous morphology
Pathology review of GNAS mutant tumors revealed a prominent villous morphology in seven of eight (88%) cases available for review. In five of these cases, the GNAS mutant cancers arose in a contiguous villous adenoma. In two of these cases, the cancers themselves demonstrated a highly unusual villous architecture (Figure 1.4). Villous
adenomas are a subtype of adenomatous polyp, accounting for approximately 5-15% of adenomas [61]. Villous cancers, however, are currently not recognized as a specific sub-classification of CRC, though studies suggest that villous adenocarcinomas may account for approximately 9% of CRC cases [62,63]. Our results indicate that GNAS mutant tumors near exclusively arise in association with villous morphology present in either the cancer or the antecedent adenoma.

A recent study by Yamada and colleagues detected GNAS codon 201 hotspot mutations in 83% villous adenomas of the colon and rectum among Japanese patients [56]. We accordingly extended our mutation analysis to a small panel of villous, tubulovillous, and tubular adenomas to determine the frequency of GNAS mutations in a western adenoma cohort (Table 1.3). Overall, we found GNAS codon 201 mutations in 46% (6/13) of villous adenomas and in 15% (3/20) of tubulovillous adenomas, and in no tubular adenomas. Thus GNAS mutations in adenomas arise exclusively in the context of villous morphology. However, in the western population we characterized we find that villous adenomas divide molecularly into two approximately equal sized groups, one group comprised of villous adenomas that arise via the GNAS mutation pathway, and the other set of villous adenomas that arise through an alternative pathway.

1.5.3 GNAS mutant CRCs are predominantly KRAS/BRAF mutant and located in the proximal colon

Among IPMNs, GNAS mutations are frequently co-accompanied by mutations of KRAS. Among GNAS mutant colon cancers, we identified mutations in the KRAS and BRAF oncogenes in nine of the ten (90%) cases, a fraction that is significantly higher than the
overall frequency of KRAS and BRAF mutations in the tumor cohort (P < 0.0305, Fisher’s exact test). The overall frequency of KRAS and BRAF mutations in our tumor population was approximately 56% (44% KRAS, 12% BRAF), which is in agreement with previous studies [64]. When examined for anatomical location within the colon, all ten GNAS mutant samples were found to be derived from primary cancers that originated in the proximal colon, between the cecum and splenic flexure, lesions commonly referred to as right-sided (P < 0.007, Fisher’s exact test). Included in our tumor cohort were both microsatellite stable (MSS) and microsatellite unstable (MSI) colon tumors, with MSI cases arising predominantly on the right side. Because three GNAS mutant tumors were also MSI tumors, it is possible that including MSI cancers skews the association of GNAS mutations with the proximal colon. However, when all MSI cancers are excluded from the analysis, the association of GNAS positive cancers with the proximal colon remains significant (P < 0.044)

It is interesting to note that while all GNAS mutant cancers were right-sided, GNAS mutant adenomas were found throughout the colon (Table 1.4). No significant differences were observed in GNAS mutant tumors when analyzed for gender, ethnicity, clinical stage, or microsatellite status.

1.6 Conclusions

In this study we find that GNAS mutations associate with a distinct subclass of colon cancers that is typified by location in the proximal colon, by having coincident KRAS or BRAF mutation, and by association with villous morphology. Although the frequency of GNAS mutant colon cancers is 1.3%, we find these tumors constitute a distinct
molecular-pathologic subclass of colon cancer. To our knowledge, this analysis of 428 colon tumors is the most extensive analysis of these mutations that has been done. The 1.3% frequency of GNAS mutations in this cohort, is higher than that reported most recently by Idziaszczyk and colleagues in 2010, though less than initially observed in a previous smaller study by we and collaborators [57]. Intriguingly, GNAS mutant cancers are not found in the distal colon, whereas GNAS mutant villous adenomas arise throughout the colon, raising the possibility that GNAS mutant adenomas progress more rapidly to cancer in the proximal colon or are more likely to become symptomatic and detected when located in the distal colon. These questions will need to be further addressed experimentally to distinguish between these two possibilities.

1.7 Discussion and Future Directions

The findings of this study are important in that they establish a reasonable estimate of the mutation frequency of GNAS in CRC and describe an association of GNAS mutations with a unique molecular and pathologic phenotype. Though our initial observation of GNAS mutations associating with late stage disease was not validated, there are several findings with clinical relevance and potential translational aims. We found a somewhat low GNAS mutation frequency in CRC of ~2.3%, however, given the estimate of 130,000 new CRC cases in the year 2015, this would approximate 2,900 GNAS mutant CRC tumors, which is equivalent to the number of all estimated osteosarcoma cases in U.S. for the same year[1]. Thus even with low frequency mutation such as GNAS, the overall prevalence of CRC results in a substantial number of cases with rare mutations, making GNAS still an attractive candidate for targeted therapy. Functional studies of
GNAS in APC<sup>min/+</sup> mice indicate its constitutive activation promotes tumorigenesis in this model, supporting its function as an oncogene in CRC. Interestingly, a compound specifically targeting Gsa, BIM-46174, has been shown to induce apoptosis <i>in vitro</i> and arrest tumor growth in pre-clinical mouse xenograft models, providing proof-of-principle that targeted inhibition of GNAS may indeed be feasible and applicable [65,66].

The finding of coincident BRAF/KRAS mutations in a high proportion of GNAS mutant tumors is of immediate clinical importance. Therapies targeting EGFR are currently the most successful in staving off metastatic CRC progression, and resistance to such therapies is most commonly due to activating KRAS, BRAF, and PIK3CA mutations[67,68]. Therefore, simply the presence of a GNAS mutation most likely indicates resistance to EGFR inhibition and identifies a subset of the patient population in need of alternative treatment options.

The observation that GNAS mutant tumors are restricted to the proximal, right side of the colon is curious. Distinct molecular pathways between right and left sided colon cancer have been known for sometime, as right sided cancers are more likely to be MSI, BRAF mutant, and CpG island methylated phenotype (CIMP) than their left sided counterparts which are characterized molecularly by genomic instability and aneuploidy[69]. It should be noted that this distinction is not anatomically arbitrary and that there are bona fide differences between the proximal and distal colon. Developmentally, the proximal colon, stretching from the caecum, ascending colon, and the proximal 2/3<sup>rd</sup> of the transverse colon, is derived from the embryonic midgut whereas the distal 1/3<sup>rd</sup> of the transverse colon, descending and sigmoid colon, and rectum are derived from the embryonic hindgut[70]. Consequently, this results in a different vascular
supply as well as distinct physiologic differences in bile acid concentrations, composition of bacterial flora and bacterial metabolites, and different carcinogen exposures[71,72]. Though largely speculative, these distinctions are in part thought to contribute to the different molecular pathways operating in right versus left sided colon cancers. Regardless of the underlying genetic aberrations, anatomical location alone has prognostic significance as metastatic cancers of right-sided origin portend a poorer clinical outcome[73].

Adenomas of the colon are histologically subtyped by architecture and classified as tubular adenoma (TA), tubulovillous adenoma (TVA), villous adenoma (VA), sessile serrated adenoma (SSA), and traditional serrated adenoma (TSA). Not surprisingly adenomas also display a proclivity for location within the colon, with a majority of TA, TVA, and VA occurring in the distal colon and SSA occurring within the proximal colon[74]. Most adenomas are of the tubular type and the identification of villous architecture is associated with the presence of high grade dysplasia and greater risk of recurrence[75]. The association of GNAS with villous morphology and the finding that GNAS mutant villous adenomas arise throughout the colon but only GNAS mutant tumors are found in the proximal colon raises the possibility that distal lesions are more frequently discovered and removed on colonoscopy or there is a selective growth advantage for this mutation in right-sided adenomas. Interestingly, villous adenomas in the distal colon have been associated with a secretory diarrhea in cancer patients that is suspected to be the result of increased cAMP signaling, similar to the mechanism of cholera toxin induced secretory diarrhea [76]. As GNAS mutant tumors would be expected to have increased cAMP production, this supports the notion that GNAS mutant
villous adenomas may indeed be discovered and removed more frequently in the distal colon, but does not rule out the possibility that GNAS mutations confer a growth advantage in the proximal colon.

Given the association of GNAS cancers with the proximal colon, even more curious is the finding of a villous architecture in cases of BRAF mutant tumors. Most proximal BRAF mutant tumors are thought to arise in the setting of SSA with sporadic MSI and CIMP phenotype[77]. In our three cases of concomitant GNAS-BRAF mutant cancers, two were found to be MSI and one of these associated with a villous lesion (Table 1.3). At first glance, this finding would seem to contradict prevailing wisdom. However, a study examining SSAs with neoplastic progression, that is SSAs containing areas of high grade dysplasia or carcinoma in situ, found that in the majority of these lesions the areas with neoplastic progression more often resembled a tubular or villous architecture and not the antecedent serrated morphology[78]. In light of this, it seems plausible that the BRAF-mutant villous cancer in our study arose from an SSA that progressed to a villous cancer upon acquiring the GNAS mutation. To test this hypothesis, a simple laser capture and micro-dissection experiment comparing the mutation profile of the SSA and villous carcinoma would allow us to determine the chronology of the GNAS mutation and if it arose before or during the progression to cancer. This would also further support the association of GNAS mutations with a villous architecture and its putative oncogenic function in CRC.

Follow-up studies to ours will need to be conducted to provide more statistical power to the restriction of GNAS mutant tumors to the proximal colon and the significance of these mutations on clinical outcome. A recent study using routine
genotyping for a panel of genes in CRC tumors, including GNAS, found similarly that GNAS mutations were present in about 2% of CRC tumors concurrently with KRAS or BRAF mutation and these tumors were all restricted to the proximal colon[79]. While the authors did not state any association with villous morphology, they did describe an association with a mucinous phenotype (>50% mucin content). Our pathology review did find cases of villous GNAS mutant tumors with increased mucin content (unpublished data), however this was not unexpected as villous adenocarcinomas and right-sided tumors often have a mucinous phenotype[62,80]. Recent observations that GNAS mutations are common in mucinous neoplasm of the pancreas and in appendiceal mucinous neoplasms along with functional studies showing mutant GNAS increases mucin production in pancreatic ductal cell lines support this link with a mucinous phenotype[32,81,82]. Studies examining mucin as a prognostic indicator in CRC have shown mucinous histology to be of adverse prognostic impact, most notably in MSS cancers, thus it seems logical that GNAS mutations might be prognostic in a similar manner[83].

This currently leaves us with many speculative questions regarding GNAS mutations and what their histopathologic and clinical significance may be. Do they portend a poorer prognosis? And is this prognosis due to GNAS mutations or simply because these tumors are proximal? Are they associated with both a villous and mucinous histology? What about the coincident KRAS and BRAF mutations? Of what possible prognostic significance are those in GNAS mutant tumors? With routine sequencing and genotyping of CRC tumors that is currently being performed, these questions can be addressed quite simply. It seems feasible that a retrospective study using a few hundred
GNAS mutant cases could be conducted that would allow us to make the necessary comparisons to answer the above questions. In this way we could achieve statistical power to address the issues of GNAS mutations and associations with histologic subtype; patient outcomes based on KRAF/BRAF mutations, anatomic location, and MSI/MSS status; and finally GNAS mutations as a predictor to treatment response including traditional chemotherapy as well as targeted therapeutics.
1.8 Figures and Tables

Figure 1.1 Genetic pathways frequently targeted in Colon cancer. Mutation of APC in the WNT pathway is the most common event in CRC and occurs in early in the transition to an adenoma. Activating KRAS mutations are the most frequent oncogenic aberrations. The final step in progression to an invasive cancer is inactivation of the tumor suppressor p53 and/or TGFBR1. Pathways promoting tumor growth upregulated in CRC include the pro-inflammatory COX-2 pathway and EGFR signaling.
Figure 1.2 Somatic mutation landscape of CRC. (Top) Each position in the landscape represents a unique genomic coordinate. The more frequent the mutation event in CRC at a position, the taller the resultant peak. As expected, mutations in APC, KRAS, and p53 were the most common, the rest of the landscape is dominated by hills, which represent low frequency recurring mutations in CRC. (Bottom) Comparison of the landscape in two CRC tumors. Individual yellow and white dots are mutations unique to either tumor. Blue dots show shared mutations. There are many unique mutations in these tumors and only three commonly shared mutations.
Table 1.1 Tumor molecular and clinical characteristics according to GNAS mutation status

<table>
<thead>
<tr>
<th>GNAS Status</th>
<th>GNAS WT (%)</th>
<th>GNAS Mutant (%)</th>
<th>Total</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>GNAS WT</td>
<td>418 (97.7%)</td>
<td>10 (1.3%)</td>
<td>428</td>
<td></td>
</tr>
<tr>
<td>GNAS Mutant</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>428</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Tumor Location</th>
<th>GNAS WT (%)</th>
<th>GNAS Mutant (%)</th>
<th>Total</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Left</td>
<td>129 (30.9%)</td>
<td>0</td>
<td>129</td>
<td></td>
</tr>
<tr>
<td>Right</td>
<td>195 (46.7%)</td>
<td>10 (100%)</td>
<td>205</td>
<td></td>
</tr>
<tr>
<td>N/A</td>
<td>94 (21.5%)</td>
<td>0</td>
<td>94</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>418</td>
<td>10</td>
<td>428</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>KRAS/BRAF Status</th>
<th>GNAS WT (%)</th>
<th>GNAS Mutant (%)</th>
<th>Total</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>KRAS/BRAF WT</td>
<td>182 (43.5%)</td>
<td>1 (10%)</td>
<td>183</td>
<td></td>
</tr>
<tr>
<td>KRAS/BRAF Mut</td>
<td>236 (56.5%)</td>
<td>9 (90%)</td>
<td>245</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>418</td>
<td>10</td>
<td>428</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Gender</th>
<th>GNAS WT (%)</th>
<th>GNAS Mutant (%)</th>
<th>Total</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Male</td>
<td>207 (49.5%)</td>
<td>5 (50%)</td>
<td>212</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>211 (50.5%)</td>
<td>5 (50%)</td>
<td>216</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>418</td>
<td>10</td>
<td>428</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ethnicity</th>
<th>GNAS WT (%)</th>
<th>GNAS Mutant (%)</th>
<th>Total</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Caucasian</td>
<td>337 (80.4%)</td>
<td>8 (80%)</td>
<td>345</td>
<td></td>
</tr>
<tr>
<td>African Amer</td>
<td>71 (17.2%)</td>
<td>2 (20%)</td>
<td>73</td>
<td></td>
</tr>
<tr>
<td>N/A</td>
<td>10 (1.4%)</td>
<td>0</td>
<td>10</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>418</td>
<td>10</td>
<td>428</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Stage</th>
<th>GNAS WT (%)</th>
<th>GNAS Mutant (%)</th>
<th>Total</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage I&amp;II</td>
<td>144 (34.4%)</td>
<td>3 (30%)</td>
<td>147</td>
<td></td>
</tr>
<tr>
<td>Stage 3&amp;4</td>
<td>274 (65.6%)</td>
<td>7 (70%)</td>
<td>281</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>418</td>
<td>10</td>
<td>428</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MMR status</th>
<th>GNAS WT (%)</th>
<th>GNAS Mutant (%)</th>
<th>Total</th>
<th>P-Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>MSS</td>
<td>369 (88.3%)</td>
<td>7 (70%)</td>
<td>376</td>
<td></td>
</tr>
<tr>
<td>MSI</td>
<td>38 (9.1%)</td>
<td>3 (30%)</td>
<td>41</td>
<td></td>
</tr>
<tr>
<td>N/A</td>
<td>11 (1.6%)</td>
<td>0</td>
<td>11</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>418</td>
<td>10</td>
<td>428</td>
<td></td>
</tr>
</tbody>
</table>

Abbreviations: MMR, Mismatch Repair; MSS, Microsatellite Stable; MSI, Microsatellite instability
**Figure 1.3: GNAS Codon 201 Mutations.** Representative chromatograms (left) and pyrograms (right) of GNAS wild-type, GNAS R201C, and GNAS R201H mutations detected in colon cancer. Arrows in chromatograms show mutant, heterozygous peaks. Boxed peaks in pyrograms highlight mutant alleles. Percentages indicate allele frequencies calculated from pyrogram peak intensities.
**Table 1.2:** Clinical and molecular characteristics of GNAS mutant tumors

<table>
<thead>
<tr>
<th>ID</th>
<th>Stage</th>
<th>GNAS Result</th>
<th>Kras Result</th>
<th>Braf Result</th>
<th>Location</th>
<th>MMR status</th>
<th>Gender</th>
<th>Ethnicity</th>
<th>Morphology</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>II</td>
<td>R201H</td>
<td>G12D</td>
<td>WT</td>
<td>right</td>
<td>MSS</td>
<td>Female</td>
<td>Caucasian</td>
<td>Villous</td>
</tr>
<tr>
<td>2</td>
<td>II</td>
<td>R201H</td>
<td>WT</td>
<td>V600E</td>
<td>right</td>
<td>MSS</td>
<td>Female</td>
<td>Caucasian</td>
<td>Villous</td>
</tr>
<tr>
<td>3</td>
<td>II</td>
<td>R201C</td>
<td>WT</td>
<td>V600E</td>
<td>right</td>
<td>MSI</td>
<td>Male</td>
<td>Caucasian</td>
<td>Villous</td>
</tr>
<tr>
<td>4</td>
<td>III</td>
<td>R201C</td>
<td>WT</td>
<td>V600E</td>
<td>right</td>
<td>MSI</td>
<td>Male</td>
<td>Caucasian</td>
<td>Signet</td>
</tr>
<tr>
<td>5</td>
<td>IV</td>
<td>R201H</td>
<td>G12V</td>
<td>WT</td>
<td>right</td>
<td>MSS</td>
<td>Female</td>
<td>Caucasian</td>
<td>Villous</td>
</tr>
<tr>
<td>6</td>
<td>IV</td>
<td>R201H</td>
<td>G12V</td>
<td>WT</td>
<td>right</td>
<td>MSS</td>
<td>Female</td>
<td>Afr Amer</td>
<td>Villous</td>
</tr>
<tr>
<td>7</td>
<td>IV</td>
<td>R201H</td>
<td>WT</td>
<td>WT</td>
<td>right</td>
<td>MSI</td>
<td>Male</td>
<td>Caucasian</td>
<td>N/A</td>
</tr>
<tr>
<td>8</td>
<td>IV</td>
<td>R201H</td>
<td>G12V</td>
<td>WT</td>
<td>right</td>
<td>MSS</td>
<td>Male</td>
<td>Caucasian</td>
<td>N/A</td>
</tr>
<tr>
<td>9</td>
<td>Met</td>
<td>R201H</td>
<td>G13D</td>
<td>WT</td>
<td>right</td>
<td>MSS</td>
<td>Female</td>
<td>Caucasian</td>
<td>Villous</td>
</tr>
<tr>
<td>10</td>
<td>Met</td>
<td>R201C</td>
<td>A146T</td>
<td>WT</td>
<td>right</td>
<td>MSS</td>
<td>Male</td>
<td>Afr Amer</td>
<td>Villous</td>
</tr>
</tbody>
</table>

Abbreviations: MMR, mismatch repair; Met, metastasis; Afr Amer, African American
Figure 1.4: GNAS tumors are associated with a villous morphology. H&E Photomicrograph of a GNAS mutant, villous adenocarcinoma. The tumor has typical villous morphology with protruding papillae containing a fibrovascular core and lined with adenomatous epithelium.
Table 1.3: GNAS mutation frequency in adenomas

<table>
<thead>
<tr>
<th>Adenoma type</th>
<th>Count</th>
<th>GNAS Mutant (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tubular</td>
<td>17</td>
<td>0</td>
</tr>
<tr>
<td>Tubulovillous</td>
<td>20</td>
<td>3 (15%)</td>
</tr>
<tr>
<td>Villous</td>
<td>13</td>
<td>6 (46%)</td>
</tr>
</tbody>
</table>

Table 1.4: Characteristics of GNAS mutant adenomas

<table>
<thead>
<tr>
<th>ID</th>
<th>GNAS status</th>
<th>Morphology</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ad-1</td>
<td>R201H</td>
<td>TVA</td>
<td>right</td>
</tr>
<tr>
<td>Ad-2</td>
<td>R201H</td>
<td>TVA</td>
<td>right</td>
</tr>
<tr>
<td>Ad-3</td>
<td>R201H</td>
<td>TVA</td>
<td>right</td>
</tr>
<tr>
<td>Ad-4</td>
<td>R201C</td>
<td>VA</td>
<td>left</td>
</tr>
<tr>
<td>Ad-5</td>
<td>R201H</td>
<td>VA</td>
<td>transverse</td>
</tr>
<tr>
<td>Ad-6</td>
<td>R201H</td>
<td>VA</td>
<td>rectum</td>
</tr>
<tr>
<td>Ad-7</td>
<td>R201H</td>
<td>VA</td>
<td>rectum</td>
</tr>
<tr>
<td>Ad-8</td>
<td>R201C</td>
<td>VA</td>
<td>right</td>
</tr>
<tr>
<td>Ad-9</td>
<td>R201C</td>
<td>VA</td>
<td>left</td>
</tr>
</tbody>
</table>

Abbreviations: TVA, tubulovillous adenoma; VA, villous adenoma
CHAPTER 2. GENETIC STUDY OF FAMILIAL BARRETT’S ESOPHAGUS AND ESOPHAGEAL ADENOCARCINOMA IDENTIFIES VSIG10L AS A CANDIDATE DISEASE SUSCEPTIBILITY GENE

2.1 Introduction

2.1.1 Study of Familial Cancer Syndromes

Family studies of hereditary cancer syndromes have been indispensable for identifying genes and pathways involved in neoplasia. Since the discovery of the first cancer susceptibility gene and tumor suppressor RB1 in hereditary retinoblastoma, studies of familial cancers have proven fruitful in providing insight into the underlying genetic and cellular mechanisms of carcinogenesis[84]. Though familial cancer susceptibility syndromes are thought to only account for a small fraction of all cancers, germ-line mutations in familial cancers often occur in genes or pathways frequently targeted in sporadic disease[85]. For example, germ-line mutations in the tumor suppressor gene APC are responsible for the syndrome familial adenomatous polyposis (FAP), characterized by the development of hundreds to thousands of adenomatous polyps in the colon of patients at an early age with almost certain progression to cancer by the age of 50[86]. While FAP itself is estimated to be responsible for only ~1% of CRC cases per year, mutations in the gene APC are found in greater than 80% of sporadic tumors[7]. Thus, identification of this critical gene in a rare syndrome was hugely relevant in understanding a pivotal mechanism acting in the majority of sporadic colon cancers.

Even in the age of next generation sequencing, family studies still offer advantages for novel gene discovery over the brute force sporadic cancer sequencing
studies being routinely performed today. Sequencing large numbers of sporadic tumors gives statistical power for identifying candidate genes, but many of the described gene mutations occur in genes of unknown function or the specific contribution of a infrequent gene mutation to the cancer phenotype has not been established, rendering biological inferences unfeasible[87]. Familial studies of cancer describing a single, causal gene variant give a direct correlation of genotype to phenotype, allow for focusing functional efforts on few candidate gene variants as opposed to many, and can assist in elucidating the functional consequences of sporadic disease variants in the same or similar genes. This approach has been applied successfully to discover important and novel disease variants in many diseases, and coupling this approach with the power of next generation sequencing has yielded much success[88,89]. Here, we will review whole exome sequencing for gene discovery in familial cancers.

2.1.2 Next Generation Sequencing Overview

Next generation sequencing technologies all share one similar feature: massively parallel sequencing of individual DNA molecules spatially separated on a flow cell (Figure 2.1)[90]. The widely used Illumina platform utilizes a synthesis by sequencing method whereby incorporated fluorescently labeled nucleotides are imaged after each extension cycle, identifying each incorporated base at a time. Upon completion of sequencing a DNA molecule, the raw sequence is used to generate what is referred to as a read, generally a sequence of 35-100bp in length. To determine the genomic location of a given read, alignment of reads to a reference genome is conducted. Once reads are aligned, identifying variants in reads that differ from the reference genome (e.g. SNPs, indels) is performed using variant calling algorithms. Following variant identification,
variant annotation is performed to obtain the position and resulting change (e.g. position of a missense variant in a gene)[91]. The functional consequences of identified variants can then be assessed *in silico* with commonly used variant prediction algorithms. The entire analysis pipeline beginning with the actual sequencing to the final variant annotation step can be completed in under a month.

2.1.3 Whole Exome Sequencing for Mutation Discovery in Familial Cancer

Next generation sequencing technology can certainly be used to interrogate the entire genome, but this approach has limited utility in familial cancer studies. Only around 10% of the human genome sequence is functionally characterized, leaving us with insufficient knowledge to make inferences regarding variation in the majority of the genome[92]. In addition to cost, there is a computational and analytical burden in attempting to annotate and sort through an excess of data. To facilitate discovery of variants underlying a disease, targeted sequencing approaches have been adapted that enrich for a desired proportion of the genome. One such approach, whole exome sequencing (WES), is the targeted enrichment of all the coding regions of the genome that has been adopted for both germ-line and somatic mutation discovery [93]. Most familial cancers follow a Mendelian pattern of inheritance and most alleles known to underlie Mendelian disorders disrupt protein-coding genes, thus WES has been widely used for novel disease gene discovery in many familial cancer syndromes[87].

The workflow for conducting WES is outlined in (Figure 2.2). The enrichment step in WES consists of hybridizing the genomic library to probes specific for only the coding regions. To begin a study of familial cancer, one first needs to identify a family or
families exhibiting a heritable cancer phenotype. Different sequencing strategies could be used from here depending on the available pedigrees (Figure 2.3). With multiple small pedigrees the idea is to sequence across multiple unrelated, affected individuals and to identify novel variants in the same gene. If a multigenerational, multiplex family has been accrued, then the strategy is to sequence distantly related affected family members and identify the novel shared variants[93]. In instances of singular cases or single small pedigrees, novel variants identified can be screened for in population controls or in other familial cases to provide more evidence that the variant is pathogenic.

Following exome sequencing of family members comes the difficult process of variant filtering. The average exome from a single individual will reveal approximately 25,000 variants, thus the challenge lies in identifying which of these variants is disease causal. A simple and obvious first step is to exclude synonymous changes and select only protein-altering variants, such as those leading to non-synonymous, frame-shift, or splice-site changes. Next one would filter out common variants by removing those known to existing in the general population, as these are unlikely to cause disease. Databases such as the 1000 genomes project, dbSNP, and the Exome Variant Server, which have compiled allele frequencies on thousands of sequenced exomes, are routinely used for this task[94]. For an autosomal dominant disorder, variants with a population allele frequency >1% is more than sufficient for filtering, as dominant disease alleles would be expected to be very rare (<0.1%)[93]. Next we would look for only those variants being shared in the affected members sequenced, as most, if not all, affected members should harbor the disease allele. Variants remaining after this step can be ranked based on the predicted impact of the variant on the protein, such as weighing non-sense and frameshift
mutations greater than non-synonymous changes. One may also use any known function of a gene to inform the ranking, such as its role in biological pathways or its association with any other known cancer phenotypes. Strategies for ranking missense variants might include assessing conservation of an amino acid residue by multiple sequence alignment and using functional protein prediction algorithms that consider both conservation and physiochemical consequences of a particular amino acid substitution[95,96].

Using these filtering schemes it is possible to narrow down the list of candidates to just a few, and perhaps one, candidate disease susceptibility variant. With a manageable number of candidates in hand, it is simple enough to perform segregation analysis by sequencing for the variant in the remaining affected and unaffected members of a family and determining if the variant is co-segregating with disease. Those variants remaining after segregation analysis can in turn be tested in functional studies using a suitable disease model to establish a mechanism for disease susceptibility. While the described filtering process has been successful for discovering novel disease genes in many instances, there are some caveats. Technical limitations of sequencing that may preclude variant detection include incomplete target capture, poor sequencing coverage, and incorrect variant annotation. More likely though are the analytic failures due to disease complexity and genetic heterogeneity. Such examples include incomplete penetrance, very rare disease alleles limited to one or a few families, and the presence of phenocopies (i.e. family members with disease but do not harbor the variant). Further, as WES only captures coding DNA, possibly disease variants in regulatory or non-coding elements will be missed.
Caveats aside, the power of WES for gene discovery is undeniable, with over 150 novel disease genes described since its inception in 2009[97]. Though there are hurdles to adopting the technology, notably establishing the bioinformatics analysis pipeline, it offers many advantages over traditional linkage mapping. WES can be used for rare cases of disease and small pedigrees, find disease variants in small regions of sharing, and detect low penetrant variants[98]. Further it bypasses some of the problems of linkage analysis such as complicated statistical modeling and the inherent problems accompanying genotyping (lack of informative markers, small shared regions etc.). Given these advantages, we selected the WES approach to identify a disease susceptibility variant in a large family pedigree with Familial Barrett’s esophagus, a hereditary cancer syndrome of Barrett’s esophagus and esophageal adenocarcinoma.

2.1.4 Familial Barrett’s Esophagus

Barrett’s esophagus (BE) is defined as the replacement of the normal stratified squamous epithelium of the esophageal mucosa with a metaplastic, columnar epithelium resembling that of the small intestine[99] (Figure 2.4). The development of BE is considered to be an adaptive response to inflammation and tissue injury due to chronic gastroesophageal reflux disease (GERD). The presence of BE is asymptomatic but is of clinical concern as it is the only known precursor lesion for esophageal adenocarcinoma (EAC), an aggressive cancer whose incidence has increased over 6-fold in the past 30 years [100,101]. Risk factors contributing to the development of BE and EAC include male gender, Caucasian ethnicity, symptomatic GERD, and obesity. The prognosis of patients with EAC is poor as the disease carries a 5-year survival rate around 17%, in part because
the majority of presenting cases (~70%) already have either regional or distant metastases. Difficulties with early detection arise from the fact that BE carries a modest risk of progression to EAC, as estimates suggest the annual risk for developing EAC in patients with BE is around 0.2%/year, making surveillance programs ineffective and costly [102]. Improving treatment of EAC has also been difficult due to marked heterogeneity in the sporadic mutation landscape between tumors, hindering the application of targeted therapeutics[103]. These obstacles necessitate an improved understanding of the biology underlying BE and EAC.

An inherited susceptibility to developing EAC is well documented and current estimates suggest that around 7% of patients presenting with either BE or EAC have a confirmed affected first or second-degree relative[104]. This familial clustering has been termed familial Barrett’s esophagus (FBE), defining the aggregation of BE and EAC as single trait[105]. As in other familial cancer syndromes, FBE demonstrates an autosomal dominant pattern of inheritance with incomplete penetrance[106]. Further supporting this inherited predisposition is the fact that individuals with EAC from multiplex families are diagnosed on average 7 years earlier than non-familial cases (56 vs. 63 years of age)[107]. This earlier age of diagnosis is another hallmark of cancer susceptibility syndromes. Given the epidemiologic support for FBE as a familial cancer syndrome and the previous successes of family studies in identifying genes contributing to carcinogenesis, we elected to pursue a whole exome sequencing study of a large FBE family to find a novel gene, or genes, underlying disease susceptibility.
2.2 Approach

We identified a multiplex, multigenerational FBE family with 14 affected members: 11 affected individuals with BE and 3 affected with EAC, thus exhibiting a strong genetic predisposition to disease development. From this family we selected 4 affected individuals for whole exome sequencing to identify a germ-line disease susceptibility variant or variants. We selected distantly related individuals to reduce the number of shared genetic variants and chose those individuals showing a strong susceptibility phenotype (i.e. young age at diagnosis, female affecteds) to decrease the likelihood of sequencing phenocopies. Following sequencing we narrowed down variants sequentially by looking for only protein-altering variants co-segregating in the four members sequenced, filtering against population polymorphisms using variant databases, and used variant prediction algorithms to identify variants predicted to be deleterious to protein function. Partial functional characterization was performed in a 3D organotypic cell culture model of esophageal maturation to ascertain biological impact candidate gene variant on disease development.

2.3 Abstract

Barrett’s esophagus and esophageal adenocarcinoma aggregate in families. Using whole-exome sequencing in a multigenerational Familial Barrett’s Esophagus (FBE) family, we identified a private, germline missense variant at a highly conserved serine residue (S631G) in an uncharacterized gene, VSIG10L, that segregated in affected members. Transfection of S631G variant into a 3D-organotypic culture model of normal squamous cells dramatically inhibited epithelial maturation compared to the wild-type. VSIG10L exhibited high expression in normal squamous esophagus with marked loss of expression
in Barrett’s-associated lesions. Taken together, we present VSIG10L as a candidate FBE susceptibility gene, with a putative role in maintaining normal esophageal homoeostasis.

2.4 Materials and Methods

Proband and Family Recruitment

The proband and family in this study were identified and recruited in an institutional review board (IRB) approved study at Case Western Reserve University and University Hospitals of Cleveland Medical Center as previously described[108]. Family members who had no prior esophagogastroduodenoscopy (EGD) were offered screening[108]. The phenotype definitions were as follows -- Long segment Barrett’s esophagus (BE) = intestinal metaplasia on biopsy plus ≥ 3 cm segment on EGD; short segment BE = intestinal metaplasia on biopsy plus >1, < 3 cm segment, esophageal adenocarcinoma (EAC) was defined as adenocarcinoma on biopsy report involving tubular esophagus. Blood lymphocytes were collected, immortalized, and banked at the Rutgers University DNA Repository (RUCDR) as a source of germ-line DNA from all family members who had screening endoscopy. Formalin fixed paraffin embedded specimens from deceased family members with EAC were obtained with consent when available.

Tissue Samples

Normal esophagus squamous, BE, and EAC tissue samples for real-time PCR analysis of VSIG10L expression were obtained from participating institutions in the Barrett’s Esophagus Translational Research Network (BETRNet). All biopsy samples were collected under an Institutional Review Board for Human Subject Investigation approved
protocol. Tissue was collected and banked at -80°C from patients with BE undergoing routine surveillance and patients with newly diagnosed EAC undergoing endoscopic procedures. Endoscopic biopsies were immediately snap frozen at bedside to preserve RNA. Biopsies from patients with non-dysplastic stable BE (NDSBE) or high grade dysplasia (HGD) were obtained preferentially from areas likely to have intestinal metaplasia or high grade dysplasia, respectively, using high definition white light and narrow band imaging guidance[109–111]. For this study, a total of 103 normal esophagus squamous, 68 NDSBE, 21 HGD, and 46 EAC (N=240 for entire cohort) were obtained for RNA extraction and subsequent real-time analysis.

Whole-exome capture and deep sequencing

Target capture, library preparation, and deep sequencing were performed by the Oklahoma Medical Research Foundation Next Generation DNA Sequencing Core Facility (Oklahoma City, OK), as previously described[112]. Target sequence enrichments were performed using the Illumina TruSeq Exome Enrichment Kit as per the manufacturer's protocols (Illumina Inc). Briefly, sample DNAs were quantified using a picogreen fluorometric assay and 3 μg of genomic DNA were randomly sheared to an average size of 300 bp using a Covaris S2 sonicator (Covaris Inc). Sonicated DNA was then end-repaired, A-tailed, and ligated with indexed paired-end Illumina adapters. Target capture was performed on DNA pooled from six indexed samples, following which the captured library was PCR amplified for 10 cycles to enrich for target genomic regions. The captured libraries were precisely quantified using a qPCR-based Kapa Biosystems Library Quantification Kit (Kapa Biosystems) on a Roche Lightcycler 480
(Roche Applied Science). Deep sequencing of the capture enriched pools was performed on an Illumina HiSeq 2000 instrument with 100 bp, paired-end reads to an average read-depth of 70× per sample.

**Read mapping, variant detection, and annotation**

Burrows-Wheeler Aligner (BWA)[113] or Short Oligonucleotide Analysis Package (SOAP)[114] algorithms were used to align individual 100-bp reads from the raw FASTQ files to the human reference genome (build hg19). Following the conversion of aligned reads in to binary Sequence Alignment/Map (BAM) format, coverage metrics of target bases were calculated using the Picard algorithm ([http://samtools.sourceforge.net](http://samtools.sourceforge.net)). On average, Picard metrics showed >80% of the target bases covered at 20× read-depth for the samples, with approximately 6% of target bases showing no coverage. Next, sequence variations (both single nucleotide and insertion/deletion) in the germline of respective samples were detected using two variant calling algorithms including, Genome Analysis Toolkit (GATK)[115], and mPILEUP[116]. Genomic variants were mapped to the human transcriptome reference database (RefSeq, build hg19) using a variant annotation tool developed in house (SLATE) that identifies variants mapping to gene coding regions and splice-sites, including their corresponding positions and codon changes within respective transcripts.

**Filtering of variants and identification of **\(\text{VSIG10L}\)** as a candidate FBE susceptibility gene**

Given a dominant inheritance pattern for the disease in this family, we identified all variants including, non-synonymous single-nucleotidie variants (SNV), insertion/deletion variants (indel), and splice-site variants in gene coding regions that were co-segregating in the four FBE31 affected individuals (III-4, III-8, IV-1, and IV-17). Next, we
eliminated variants with a minor allele frequency (MAF) ≥6% based on the public 1000 genomes database[117], the NHLBI GO Exome Sequencing Project (ESP) (URL: http://evs.gs.washington.edu/EVS/) database, and an in-house platform-matched whole-exome sequencing database derived from 106 random non-Barrett’s germline samples. The use of platform-matched in-house database additionally aided in eliminating recurrent artifacts or false positive calls seen in the FBE31 exome dataset. This resulted in a total of 4 protein-altering variants (3 missense and 1 frameshift deletion) co-segregating in the four FBE31 affected individuals. Finally, aligned reads mapping to each of the 4 variant genomic loci were manually reviewed using the Integrative Genomics Viewer[118] to confirm their true variant statuses. Of the 4 co-segregating variants identified, 3 variants mapped to genes DCHS2, LRRC43, and BCLAF1, respectively, with each variant being detected in 2%-5% of control population, based on the public and in-house germline whole-exome databases mentioned above. The remaining missense variant, mapping to the gene VSIG10L (c.1891A>A/G, p. 631S>S/G), was a totally private variant not seen in any of the public or in-house germline whole-exome databases, as well as a control cohort consisting of 126 DNA samples from symptomatic GERD patients that were endoscopically confirmed as BE negative. The impact on biological function of the 631S>G missense variant was predicted to be possibly deleterious by the protein variant analysis tools PROVEAN[119], mutationassessor[95], and PolyPhen-2[96]. Furthermore, a homolog in this gene family, VSIG1, has been shown to play an important role in the proper differentiation of gastric epithelium in mice[120]. Accordingly, VSIG10L was nominated as a candidate FBE susceptibility gene.
Sanger sequencing of the *VSIG10L* variant in FBE31 family members and GERD controls

Amplification of the variant encompassing region in *VSIG10L* from patient germ-line DNA was performed using the FastStart Taq DNA polymerase kit (Roche) and the following primers: Forward 5’-GACGTGGACTTCAGGGTTTT-3’ and Reverse 5’-ACTGGAGGCCTCTGGTTGT-3’. For amplification from FFPE the following primer set was used: Forward 5’-AGGCCTCTGGTTGTCCCCCA-3’ and Reverse 5’-TCCAGGTCCCAATCCAGGCTG-3’. The 5’ end of the forward and reverse primers included the forward M13 primer sequence 5’-GTAAAACGACGGCCAGT-3’ and reverse M13 primer sequence 5’-CAGGAAACAGCTATGAC-3’, respectively. All PCR reactions used 20ngs of template DNA. Cycling conditions included an initial denaturing step at 95°C for 5 min followed by 36 cycles 95°C for 30 s, 60°C for 30 s, 72°C for 45 s and a final elongation step of 72°C for 7 min.

Real-time PCR

Esophageal biopsy samples were processed for RNA extraction using the mirVana™ RNA kit according to manufacturer’s protocol (Life Technologies). cDNA conversion was carried out using the Superscript III reverse transcriptase with 1µg total input RNA in a 40µl reaction. Real-time PCR measurement of VSIG10L was performed using the human VSIG10L TaqMan Probe/Primer kit Hs01039199_g1 from Life Technologies. Beta-2 microglobulin (B2M) TaqMan Probe/Primer kit (4333766) from Life Technologies was used as the endogenous control. A 20-µl-reaction mix containing a 1:40 dilution of primer/probe and 2.5ng of cDNA in 1x IQ Supermix from Bio-Rad was detected in an iCycler optical module (Bio-Rad). Thermal cycling for both assays was
initiated at 95°C for 10 min, followed by 50 cycles of 95°C for 15 sec and 60°C for 1 min. All samples were assayed in triplicate for both primer/probe sets. The relative expression of VSIG10L transcript in each sample was determined by subtraction of B2M endogenous control Ct average from VSIG10L Ct average (ΔCt). VSIG10L expression in each sample was then expressed as fold-change relative to the average VSIG10L ΔCt value across the entire sample cohort (N=240).

Generation of wild-type and mutant VSIG10L expression vectors

Cloning of the VSIG10L transcript was carried out using pooled RNA isolated from normal squamous esophagus biopsies as template. Conversion to cDNA was performed using 2µg of RNA as input and Superscript III reverse transcriptase (Life Technologies). PCR amplification of the VSIG10L transcript was performed using Platinum® Pfx DNA polymerase (Life Technologies) with forward primer 5'-CACCATGGACAACCCACAGGCTCT-3' and reverse primer 5'-CACCTGTGTGGCTGCGCAGA-3'. PCR products were TA cloned into the Gateway® entry vector pCR®8/GW/TOPO®TA (Life Technologies). Following entry vector cloning, site directed mutagenesis was performed using QuikChange Lightning kit (Agilent) to introduce the germline FBE (CCDS 1891A>G; AA 631S>G) and the somatic (CCDS 2305G>A; AA 769G>S) variants. Primer sequences for site directed were as follows: FBE sense: 5'-CGCCTGGCGCTCGGTCAAGATGGGC-3'; FBE antisense: 5'-GCCCCATCTTGACCGAGCCGCAAGCC-3'; Somatic sense: 5'-GTCTACGAGCCAGCCAGCC-3'; Somatic antisense: 5'-TCAACGTGCGCTGCGCAGGCTGACG-3'. Wild type and mutant constructs were then subcloned into the Gateway® pLenti6.2/V5-DEST™ vector (Life Technologies)
according to the manufacturer’s protocol.

**Cell Culture and Infection**

Primary human esophageal keratinocytes, designated as EPC2, were established as described previously[121]. Cells were maintained at 37°C and 5% CO₂ using keratinocyte-SFM medium (KSFM; Life Technologies) supplemented with 40 µg/mL bovine pituitary extract (Life Technologies), 1.0 ng/mL EGF (Life Technologies), 100 U/mL penicillin, and 100 µg/mL streptomycin (Life Technologies). Infection of EPC2 cells with pLenti6.2-VSIG10l, pLenti6.2-VSIG10L-S631G, and pLenti-6.2-VSIG10L-G769S was performed as previously described[122]. Cells were passaged 48 hours after infection and selected with Blasticidin for seven days.

**Organotypic Culture**

Organotypic culture was performed as previously described[123]. Briefly, a collagen/Matrigel matrix, containing 76.7% bovine tendon acid-extracted collagen (Organogenesis), Matrigel Matrix (BD Bioscience), 1× minimal essential medium with Earle's salts (BioWhittaker), 1.68 mM L-glutamine (Cellgro), 10% fetal bovine serum (HyClone), 0.15% sodium bicarbonate (BioWhittaker) was mixed with 7.5×10⁴ human fetal esophageal fibroblasts. Following 7 days, 5×10⁵ human esophageal keratinocytes were seeded on top of the matrices. Cultures were fed with Epidermalization I medium for 2 days, which is a 3:1 mixture DMEM (JRH Biosciences)/Ham's F-12 (Life Technologies) supplemented with 4 mM L-glutamine, 0.5 µg/mL hydrocortisone, 0.1 mM O-phosphorylethanolamine, 20 pM triiodothyronine, 0.18 mM adenine, 1.88 mM CaCl₂, 4 pM progesterone (Sigma); 10 µg/mL insulin, 10 µg/mL transferrin, 5 mM
ethanolamine, 10 ng/mL selenium (ITES) (BioWhittaker), and 0.1% chelated newborn calf serum (Hyclone). For the next 2 days, cultures were fed with Epidermalization II medium, which is identical to Epidermalization I medium except that it contains 0.1% unchelated newborn calf serum. Then, cultures were raised to an air–liquid interface and cultured for 4 days in Epidermalization III medium, which contains the same growth supplements as Epidermalization I and II except that no progesterone is added and 2% newborn calf serum is used. Cultures were then harvested by fixing in neutral buffered formalin and later were paraffin-embedded.

**Immunohistochemistry**

Paraffin sections were dewaxed with xylene and rehydrated in graded ethanols. Antigen retrieval was performed by incubation in Antigen Unmasking Solution (Vector Laboratories) for 20 minutes at 98°C. Endogenous peroxidase activity was quenched using Peroxidazed and slides blocked in Background Sniper (BioCare Medical). Slides were incubated with anti-V5 primary antibody (Abcam, ab27671) for 1 hour at room temperature and detected with MACH 4 Universal HRP-Polymer (BioCare Medical). Signal was developed with the Betazoid DAB Chromogen kit (BioCare Medical).

**VSIG10L Protein Domain Prediction and Alignment**

The human VSIG10L protein sequence was obtained from the UniProt protein database (http://www.uniprot.org/)[124]. Of the two listed isoforms, we opted for the longest isoform (UniProt ID Q86VR7-2, isoform 2), which differs from isoform 1 by the presence of 15 additional amino acids at the C-terminal end. Importantly, the C-terminal end of isoform 2 is conserved among mammalian species and was the only isoform detected in cloning experiments. A multiple sequence alignment of human VSIG10L
protein with 11 mammalian and 5 non-mammalian orthologs was obtained using ClustalW[125] (Figure 1B). Conserved domains of human VSIG10L were predicted by InterPro[126]. Orthologous protein sequences of VSIG10L were obtained from the National Center for Biotechnology database (http://www.ncbi.nlm.nih.gov/) and their accession numbers are as follows: *Gorilla gorilla*, XP_004061329.1; *Macaca mulatta*, XP_001114712.1; *Mus musculus*, NP_001277245.1; *Rattus norvegicus*, XP_218632.4; *Canis familiaris*, XP_005616300.1, *Pan troglodytes*, XP_009434420.1; *Pteropus vampyrus*, XP_011382886.1; *Felis catus*, XP_003997612.1; *Camelus dromedarius*, XP_010985179.1; *Papio anubis*, XP_003916046.2; *Fukomys damarensis*, XP_010613981.1; *Cynoglossus semilaevis*, XP_008321218.1; *Chrysemys picta bellii*, XP_005293984.2; *Notothenia coriiceps*, XP_010791065.1; *Anolis carolinensis*, XP_008115530.1; *Danio rerio*, XP_009290965.1

2.5 Results

We previously identified a large multiplex, multigenerational FBE family, FBE family 31, exhibiting an autosomal dominant pattern of inheritance with 14 known affected members: 3 individuals with EAC and 11 individuals with BE (Figure 2.5A)[108]. Given the large number of affected individuals and young average age of cancer diagnosis (44.5 years), FBE family 31 suggested a strong genetic susceptibility to disease and was thus chosen for identification of germline disease determinants. To this end, we selected 4 individuals from this family for whole-exome sequencing (WES) (Figure 2.6). Germline DNA was not available from family members who developed EAC, we therefore selected individuals for WES based on clinical criteria suggestive of a genetic predisposition to
disease including: 1 male obligate carrier, 2 males with long segment BE at young age, and 1 female with long segment BE (BE is much less common in women) (Table 2.1)[127].

Whole exome sequencing and segregation analysis revealed only 4 candidate protein-altering germline variants co-segregating in the four affected individuals (Figure 2.6). Of these 4 shared germline variants, only the missense variant (S631G) in the gene VSIG10L was found to be completely private, i.e., not present in either the 1000 genomes database, the NHLBI GO exome sequencing project, an in-house platform-matched WES database of 106 non-Barrett’s germline samples, as well as a control cohort consisting of 126 DNA samples from symptomatic GERD patients that were endoscopically confirmed as BE negative. This strongly suggested the VSIG10L S631G variant as a novel candidate BE/EAC susceptibility allele (see Materials and Methods).

Sanger sequencing of VSIG10L in available DNA from FBE-31 family members revealed 8 of 10 affected individuals (including the 2 EAC cases), and 3 of 9 unaffected individuals as carrying the variant (Figure 2.5A, Figure 2.7). Two of the affected members with BE did not carry the S631G variant (Table 2.1). The finding of these 2 phenocopies likely reflects both the size of the FBE-31 pedigree and the ascertainment bias in identifying families with multiple BE cases. In the three instances of unaffected S631G carriers, two are young females, a group who are protected from developing BE, and the remaining carrier is a male who has developed erosive esophagitis (Table 2.1). Thus the clinical course in these individuals over time remains to be seen.

The VSIG10L gene resides on 19q12.41 and its function is unknown. The predicted product is an 881 amino acid membrane-bound protein consisting of two conserved Ig-
like domains and Ig-like folds, a single-pass trans-membrane domain, and a small cytoplasmic domain (Figure 2.5B). The described S631G variant discovered in the FBE 31 family occurs within one of the conserved Ig-like folds, and multiple sequence alignment revealed the mutated serine residue to be highly conserved in mammalian and non-mammalian species (Figure 2.5B). Furthermore, the S631G variant is predicted to be functionally deleterious by three prediction algorithms (see Materials and Methods).

The VSIG10L transcript has been found as highly expressed in the esophagus and to a lesser extent in the skin and salivary glands (http://www.ebi.ac.uk/gxa/home). Using real time PCR analysis, we confirmed that in our samples VSIG10L transcript shows high expression in the normal esophageal squamous tissue. Moreover, we additionally found that VSIG10L expression is largely lost in metaplasia, dysplasia, and adenocarcinoma lesions (Figure 2.5C). This pattern of tissue specific expression and then loss in disease suggests a potential role of VSIG10L in the normal differentiation of the squamous esophagus and in suppressing development of BE and / or EAC.

To further test for a potential functional role of VSIG10L in the esophageal epithelium, we introduced either wild type or S631G mutant VSIG10L into immortalized esophageal keratinocytes that were then tested in a 3D-organotypic cell culture (OTC) model that recapitulates the maturation of normal stratified squamous epithelium of the esophagus[128]. Cells receiving wild type VSIG10L underwent normal differentiation into keratinocytes that populated a stratified squamous epithelium (Figure 2.8). In marked contrast, cells receiving the S631G variant underwent a dramatic change in histology characterized by a massive disruption of the cellular organization of the squamous epithelium described as dysmaturation (Figure 2.8). A V5 epitope tag at the
C-terminus of both the wild-type and S613G VSIG10L constructs was equally well detected in both sets of target cells (Figure 2.8). Immunofluorescence targeted against the V5 epitope displayed robust membrane signal in the VSIG10L wild type cells while the S631G exhibited a more disordered pattern of expression (Figure 2.8). Our observation that a mutation in VSIG10L disrupts maturation of the normal esophageal squamous epithelium is further supported by recent findings that a homologous gene, VSIG1, is a cell adhesion protein whose disruption in VSIG1 wild-type and null chimeric mice disrupts epithelial differentiation in the stomach, leading to a “reverse Barrett’s” pattern of replacement of the normal glandular gastric epithilum with squamous cells [120].

To further investigate VSIG10L, we sequenced this gene in germline DNA from probands in another 35 FBE families; however no other VSIG10L private variants were identified. We therefore additionally examined VSIG10L for somatic mutations in a set of 19 esophageal adenocarcinomas. In this experiment, we identified one new somatic missense mutation, G769S, which was also predicted to be possibly deleterious. Functional testing of this variant in the 3D-OTC model showed it also inhibited squamous maturation (Figure 2.9), suggesting the VSIG10L G769S mutation likely played functional role in the development of this cancer.

2.6 Conclusions

In overview, VSIG10L shows high-level tissue specific expression in the esophagus. Germline mutation in a highly conserved residue of VSIG10L segregates together with a phenotype of familial Barrett’s esophagus and esophageal adenocarcinomas and this mutation induces dysmaturation of the squamous esophagus in functional testing in a cell
culture model. Moreover, a somatic $VSIG10L$ mutation identified in a sporadic esophageal adenocarcinoma also shows similar functional properties. Taken together, these data indicate a putative role for $VSIG10L$ in esophageal squamous maturation and/or maintenance, and support that mutation of $VSIG10L$ in FBE family 31 is likely a genetic predisposition to the development of BE and EAC in this family. We hypothesize that $VSIG10L$ functions in maturation and/or maintenance of the normal stratified squamous esophageal epithelium, and that mucosal disruption associated with the S613G VSIG10L mutation increases vulnerability to injury from reflux and other environmental insults. In overall pattern, findings for $VSIG10L$ are similar to findings of the role of $PALB2$ in pancreatic cancer, in which rare germline mutations are clearly tied to disease development in kindreds with familial pancreatic cancers, but where germline mutations in kindreds and somatic mutations in sporadic pancreas cancers are both rare events[129,130].

In summary, we describe a private germ-line variant in the gene $VSIG10L$ associated with a dominant susceptibility to BE and EAC in a large FBE family. This report is the first description of a functional, causal variant in familial Barrett’s esophagus and esophageal adenocarcinoma. Future investigations aimed at addressing $VSIG10L$ function and downstream effects of the S631G mutation should reveal as of yet uncharacterized pathways operating in esophageal maturation and susceptibility to BE and EAC.

2.7 Discussion and Future Directions

The results of our investigation illustrate the value of family studies for identification of cancer susceptibility genes and the utility of WES in this approach. The discovery of the private S631G variant in the gene $VSIG10L$ would have likely been undetected in linkage
mapping due to incomplete segregation, the presence of phenocopies in our family, and this being a rare variant, making comparison studies in other FBE families unfeasible. Using WES, we were able to circumvent these problems and build our argument for \textit{VSIG10L} as a candidate FBE susceptibility gene based on the absence of this variant in the population (i.e. it is a completely private variant), the loss of \textit{VSIG10L} expression in BE associated lesions, and functional analysis of the variant in a model of esophageal development.

Clearly there are questions arising from the observations of incomplete disease segregation and how a mutation affecting epithelial maturation may predispose to BE and EAC. The first point to address is the problem of incomplete segregation and incomplete penetrance. Of course the genetic argument would be strongest in the situation where all family members harboring the variant had disease and every unaffected was \textit{VSIG10L} wild-type. This is an inherent problem in familial cancer syndromes though, as the presence of a germ-line mutation confers only an increased risk of cancer that rises with age, as life-time risks of known susceptibility genes in other cancers range from as low as 10\% to as high as 100\%\cite{131}. Further confounding in familial cancer syndromes are specific gender biases, with many cancers exhibiting a proclivity to develop in one sex over the other\cite{132}. These observations may explain the incomplete penetrance seen in our study. In our family, we have 3 individuals with the S631G \textit{VSIG10L} variant that are disease free: 1 male and two females (Table 2.1). In the case of the male unaffected, he is 54 years of age and has developed erosive esophagitis, a known risk factor and precursor lesion to BE\cite{127}. In this case, it seems plausible that progression to the disease phenotype is emerging or has been altered by medical intervention. For the two female
carriers, both are around 40 years of age and we might expect that they develop disease at a much later age or possibly not at all. The reason being is that BE and EAC largely affect males, with a 3:1 male to female ratio for BE and 6:1 for EAC[133]. Further clinical monitoring will be needed to confirm the disease course in these individuals.

The converse of incomplete penetrance are phenocopies: individuals with sporadic disease that do not harbor the variant. In our family we found two instances of BE phenocopies as 2/10 affecteds were not carriers of the S631G variant (Table 2.1). Some of this can be explained by ascertainment bias in identifying FBE families but two other issues to address here are defining the phenotype and sporadic disease incidence. Individuals with EAC undoubtedly represent the strongest phenotype, as its incidence in the population is much lower compared to BE. The obverse is that BE is more prevalent in FBE families yet the higher incidence in the population can lead to inadvertently identifying sporadic cases in families. Had we defined our phenotype as EAC, we would have only 2 affected individuals in the family for segregation analysis, which diminishes the genetic argument. Given that almost all EAC develop from a BE lesion, we chose to treat this as a single trait, but by doing so, we introduced a greater chance of encountering phenocopies. The 2 BE phenocopies are consistent with what we would expect to find by chance. Accrued family members in this study had an upper endoscopy for symptomatic GERD. Clinical studies have shown that the incidence of BE in GERD patients is around 10-15%, thus the finding of 2/10 BE (20%) phenocopies is reasonable and in accord with previous studies[134].

From our experiments using the organotypic cell culture (OTC) model we found that overexpression of the S631G VSIG10L variant disrupted the development of the
stratified squamous epithelium and led to epithelial dysmaturation (Figure 2.8). How this may predispose to BE, and subsequently EAC, is rational in the context of BE pathogenesis. The transition from normal squamous esophagus to BE then EAC is thought to occur in a step-wise progression referred to as the metaplasia-dysplasia-adenocarcinoma sequence [135]. In this model, chronic GERD initially leads to mucosal injury and esophagitis. This transition to esophagitis is associated with the presence of elevated proinflammatory cytokines, notably IL-1B and IL-8, and an inflammatory infiltrate of T-lymphocytes[136,137]. Inflammation coupled with the physical damage caused by refluxate in turn erodes the normal stratified squamous epithelium lining the esophagus, leaving denuded foci to be repopulated by regenerating epithelial cells originating at the basal layer[138]. It was previously thought that during this step of repair Barrett’s metaplasia may arise by transdifferentiation of esophageal progenitor cells, however more recent evidence suggest that stem cells residing at the gastroesophageal junction or within the gastric cardia may migrate upward and repopulate the damaged esophagus[139–141]. These models suggest that the regeneration of the damaged esophageal mucosa may represent a competition between the native progenitor cells of the esophagus and those of a foreign lineage. In light of these observations, it seems feasible that epithelial dysmaturation caused by the S631G VSIG10L variant leads to reduced epithelial barrier function, predisposing to recurrent bouts of mucosal injury by reflux that may then be repopulated by a foreign progenitor cell giving rise to Barrett’s metaplasia.

Further precedence for a putative role of VSIG10L in BE and EAC pathogenesis comes from genome-wide association studies (GWAS) of population risk variants. These
studies have found a number of polymorphisms in genes associated with thoracic and esophageal development, notably near loci encoding the transcription factors FOXP1, BARX1, and TBX5[142,143]. The implication of population risk variants for BE and EAC in other genes contributing to esophageal development imparts added plausibility onto our findings.

Progression from Barrett’s to dysplasia and EAC is likely driven by environmental insult, as a unique mutation profile in EAC comprised of a high frequency of A>C transversions is consistent with mutation by oxidative damage [144]. Bile acids, one of the constituents of refluxate in GERD patients, are thought to promote BE and EAC through their ability to operate as aberrant signaling molecules that promote cytokine production and inflammation in the esophageal epithelia and by also inducing DNA damage[145,146]. This is further supported by evidence in an inflammatory mouse model of BE and EAC development whereby addition of bile acids to the drinking water of transgenic mice overexpressing IL-1B in the esophagus hastens progression to dysplasia and cancer[141]. Therefore we speculate that in individuals harboring the S631G VSIG10L variant there is accelerated progression to BE and in turn earlier exposure of the metaplastic epithlem to mutagens in GERD or exogenous carcinogens. The development of BE at a younger age and the likelihood of disease progression to EAC are therefore probably related to an individuals GERD severity and other environmental risk factors.

Future experiments elucidating the function of VSIG10L will be needed to address its role in epithelial maturation and downstream signaling effects. Experiments to isolate RNA from OTC cultures overexpressing the S631G variant for global gene expression
profiling are currently underway and may yield insight into affected signaling pathways. Other assays assessing its function in development could include using CRISPR/Cas to knockout \textit{VSIG10L} in esophageal keratinocytes for use in the OTC model to determine if loss of endogenous VSIG10L expression induces a similar morphology to the S631G variant or completely inhibits growth. Though a commercial antibody is not available, localization experiments in human esophageal biopsies need to be performed to assess whether VSIG10L localizes to adherens or tight junctions to determine a possible role in adhesion. Similar localization experiments of the S631G VSIG10L variant in the OTC model might also be useful in discerning if the mutation affects protein maturation and proper trafficking to the membrane. Lastly, an ambitious but not unprecedented experiment may be to use the CRISPR/Cas system for targeted somatic mutation/knockout of \textit{VSIG10L} in the esophagus of a BE mouse model to determine if this can accelerate progression to either BE, dysplasia and/or EAC[147].
2.8 Figures and Tables


**Figure 2.1 Schematic of next generation sequencing.** Single stranded DNA molecules are first attached to a flow cell. An amplification step is performed leaving double stranded bridges. The bridges are denatured leaving multiple single stranded templates for sequencing. The single stranded clusters are sequenced one base at a time with labeled fluorescent nucleotides and imaged after each elongation step. Imaging is conducted at the end of each synthesis cycle to determine the newly incorporated base. This is repeated until the length of the template has been sequenced.
Figure 2.2 Workflow schematic of whole exome library preparation. Isolated genomic DNA is sheared, adapters ligated, and coding regions enriched for by hybridizing complementary probes.
Figure 2.3 Approaches for familial sequencing studies. (A) Small pedigrees with one or few affected individuals can be used for WES by selecting multiple small families and trying to find unique germ-line variants in the unrelated individuals. (B) In larger family pedigrees, selecting individuals from different generations who are distantly related will reduce genomic sharing. Candidate disease susceptibility variants can be further narrowed down based on segregation analysis.

Figure 2.4. Illustration of anatomic and histologic features of Barrett’s esophagus.

The extension of salmon colored mucosa beyond the gastroesophageal junction found on endoscopy is diagnostic of BE. *(Top Left)* Biopsy from upper white dot. Histologically BE typically shows intestinal metaplasia with columnar, glandular epithelium adjacent to normal squamous mucosa. *(Bottom Left)* A biopsy from the lower white dot at the gastroesophageal junction demonstrating cardia mucosa that can sometimes be seen in BE lesions.
Figure 2.5. (A) Pedigree of FBE Family 31. Generations are indicated by roman numerals at left and individuals within a generation are numbered. Asterisks denote individuals from which blood samples or FFPE DNA was obtained. The proband is indicated by a gray arrowhead (individual IV-1). Clinical characteristics and affectation status are specified in Table 2.1. Exome sequencing and variant filtering work flow is outlined in Figure 2.6. (B) Map of VSIG10L predicted protein domains (Top) and multiple sequence alignment of VSIG10L orthologs (Bottom). The S631G variant discovered in FBE family 31 occurs within one of the predicted Ig-like folds as indicated by the dashed arrow on the map. The multiple sequence alignment displays a stretch of amino acids within this Ig-like fold that encompasses the mutated serine residue (arrowheads), which is highly conserved in mammalian and non-mammalian species. (C) Quantitative real-time PCR of VSIG10L transcript in normal esophageal squamous tissue (SQ), non-dysplastic Barrett’s esophagus (NDBE), high grade dysplasia (HGD), and esophageal adenocarcinoma (EAC). Expression level of VSIG10L in each of the SQ, NDBE, HGD, and EAC are shown as linear fold-change relative to mean VSIG10L expression of the entire cohort (N=240). *** represent P<0.001
Figure 2.5 VSIG10L segregation, multiple sequence alignment, and expression in BE associated lesions
Table 2.1. Clinical characteristics of members from FBE family 31 tested for VSIG10L S631G germ-line variant.

<table>
<thead>
<tr>
<th>Individual</th>
<th>Gender</th>
<th>Age at Diagnosis</th>
<th>Affectation Status</th>
<th>BE Length</th>
<th>VSIG10L Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>III-2</td>
<td>Male</td>
<td>50</td>
<td>EAC</td>
<td>-</td>
<td>S631G</td>
</tr>
<tr>
<td>III-4*</td>
<td>Male</td>
<td>65</td>
<td>BE</td>
<td>C6M6</td>
<td>S631G</td>
</tr>
<tr>
<td>III-8*</td>
<td>Female</td>
<td>75</td>
<td>BE</td>
<td>C5M5</td>
<td>S631G</td>
</tr>
<tr>
<td>IV-1*</td>
<td>Male</td>
<td>40</td>
<td>BE</td>
<td>C9M9</td>
<td>S631G</td>
</tr>
<tr>
<td>IV-6</td>
<td>Male</td>
<td>39</td>
<td>EAC</td>
<td>-</td>
<td>S631G</td>
</tr>
<tr>
<td>IV-8</td>
<td>Male</td>
<td>56</td>
<td>Esophageal Erosive</td>
<td>-</td>
<td>S631G</td>
</tr>
<tr>
<td>IV-12</td>
<td>Male</td>
<td>52</td>
<td>BE</td>
<td>C3M4</td>
<td>S631G</td>
</tr>
<tr>
<td>IV-13</td>
<td>Female</td>
<td>44</td>
<td>Unaffected</td>
<td>-</td>
<td>S631G</td>
</tr>
<tr>
<td>IV-14</td>
<td>Female</td>
<td>40</td>
<td>Unaffected</td>
<td>-</td>
<td>S631G</td>
</tr>
<tr>
<td>IV-15</td>
<td>Male</td>
<td>42</td>
<td>SSBE</td>
<td>C1M2</td>
<td>S631G</td>
</tr>
<tr>
<td>IV-17*</td>
<td>Male</td>
<td>40</td>
<td>BE</td>
<td>C4M4</td>
<td>S631G</td>
</tr>
<tr>
<td>III-6</td>
<td>Male</td>
<td>50</td>
<td>BE</td>
<td>C6M6</td>
<td>WT</td>
</tr>
<tr>
<td>III-13</td>
<td>Female</td>
<td>79</td>
<td>Unaffected</td>
<td>-</td>
<td>WT</td>
</tr>
<tr>
<td>IV-2</td>
<td>Male</td>
<td>51</td>
<td>Unaffected</td>
<td>-</td>
<td>WT</td>
</tr>
<tr>
<td>IV-3</td>
<td>Male</td>
<td>41</td>
<td>SSBE</td>
<td>C1M2</td>
<td>WT</td>
</tr>
<tr>
<td>IV-7</td>
<td>Female</td>
<td>47</td>
<td>Unaffected</td>
<td>-</td>
<td>WT</td>
</tr>
<tr>
<td>V-1</td>
<td>Male</td>
<td>22</td>
<td>Unaffected</td>
<td>-</td>
<td>WT</td>
</tr>
<tr>
<td>V-4</td>
<td>Female</td>
<td>29</td>
<td>Unaffected</td>
<td>-</td>
<td>WT</td>
</tr>
<tr>
<td>V-6</td>
<td>Female</td>
<td>22</td>
<td>Unaffected</td>
<td>-</td>
<td>WT</td>
</tr>
</tbody>
</table>

BE, Barrett’s esophagus; EAC, esophageal adenocarcinoma; SSBE, short segment Barrett’s esophagus; WT, wild type; C#M# denote circumferential and maximal extent of Barrett’s lesion measured in centimeters, respectively. * indicates individuals selected for exome sequencing.
Figure 2.6. Schematic of variant filtering workflow. The diagram shows the sequence of steps for identification and elimination of germ-line variants discovered by whole exome sequencing in the four individuals (III-1, III-8, IV-1, and IV-17). Ultimately 4 protein altering variants were found to be co-segregating in all four individuals sequenced, of which only the c.1891A>A/G (S631G) missense variant in VSIG10L was completely private and not seen in any of the whole-exome databases and in-house controls referenced (see Materials and Methods for more detail).
Figure 2.6. Schematic of variant filtering workflow
Figure 2.7 Representative chromatograms of Sanger sequencing. Wild-type *VSIG10L* in an unaffected individual (*Top*), and the c.1891A>G (S631G) germline *VSIG10L* variant in an affected member from FBE family 31 (*Bottom*). Arrows point to the c.1891A nucleotide of interest.
Figure 2.8 Photomicrographs of organotypic cell culture model expressing VSIG10L wild-type (*Left*) and VSIG10L S631G variant (*Right*). H&E analysis of wild-type VSIG10L expressed in esophageal keratinocytes shows a normal stratified squamous morphology having spherical nuclei in the basal cells that become more elongated with a parallel orientation to the surface as they mature apically. Expression of the mutant S631G VSIG10L variant (*Right*) induces a marked change in morphology characterized by epithelial dysmaturation having irregular shaped nuclei extending to the surface with loss of polarity and hyperkeratosis. Immunohistochemistry for the V5 tagged-epitope (*Middle*) shows diffuse expression of V5-tagged VSIG10L protein. Immunofluorescence targeted against the V5 epitope (*Bottom*) shows a membranous pattern of expression in the wild-type expressing cells that becomes more disordered upon introduction of the S631G variant.
Figure 2.8. Photomicrographs and immunofluorescence of organotypic cell culture model expressing WT and S631G variant
Figure 2.9. Photomicrographs of organotypic cell culture model expressing VSIG10L G769S somatic variant. The H&E stain shows disrupted epithelial maturation. Immunohistochemistry against the V5 epitope confirms expression.
CHAPTER 3: INTERROGATING THE COLON CANCER METHLYOME FOR BIOMARKER DISCOVERY

3.1 Introduction

3.1.1 Epigenetics Overview

It is well established that in addition to acquired somatic mutation, aberrant epigenetic events are key contributors to the sequential transformation of the colon epithelium into an invasive adenocarcinoma[148]. Epigenetics broadly refers to heritable variability in gene expression without any modification to the underlying genetic code[149]. These changes can be exerted primarily through three main mechanisms: DNA methylation, post-translational histone modifications, and noncoding RNA mediated gene expression. While deviations in all three mechanisms are known to contribute to tumorigenesis, aberrant DNA methylation is undoubtedly the best characterized in CRC and other cancers. Here we will primarily focus on DNA methylation changes and how they can be exploited as biomarkers for clinical monitoring of disease.

3.1.2 DNA Methylation

In vertebrates, DNA methylation occurs primarily at CpG dinucleotides by the enzymatic addition of a methyl group to the 5’-position of cytosine to produce 5-methyl-cytosine. The methylation state is regulated by DNA methyl transferases (DNMTs) which act to either maintain the methylation state during DNA replication (maintenance methylation) or initiate a new methylation events (de novo methylation)[150]. In normal cells, the majority of CpGs are methylated with the exception of CpG rich regions known as CpG islands (CGIs), defined as 200-1000bp regions with greater than 50% CpG content and an
observed to expected CpG content >0.6[151]. There is a strong correlation between CGIs and transcription, as ~60% of CGIs are in annotated promoter regions and current evidence suggests even CGIs lying intragenically may function as alternative gene promoters or in transcription of non-coding RNAs[152,153]. The conventional function of CGI methylation is to repress transcription by either direct inhibition of transcription factor binding or locking in a repressed state such as that in X-chromosome inactivation. Other biological phenomena where CGI methylation serves an important function include transposon silencing, genomic imprinting, and repression of certain germ cell genes in somatic tissues[154]. Simply stated, in normal somatic cells, most CpGs are methylated and mainly function in genome stabilization and/or imprinting with the exception of CGIs in gene promoter regions, which are usually protected from methylation.

In CRC and many other cancers, DNA methylation patterns become appreciably distorted and adopt an almost universal pattern of global hypomethylation with concomitant hypermethylation of promoter associated CGIs. Nearly three decades ago excessive loss of intergenic CpG methylation was the first epigenetic alteration to be described in colon cancer and the phenomenon of global hypomethylation has since been extended to most other cancers[155,156]. Loss of CpG methylation may have important consequences in cancer such as possible derepression of transposable elements that upon activation can contribute to genomic instability[157]. Accompanying global hypomethylation in cancer cells is simultaneous hypermethylation of CGIs in gene promoter regions. Genome-wide methylation profiling has shown that hypermethylation occurs in about 1%-10% of CGIs in individual tumors[158]. Methylation of CGIs in gene promoter regions is a frequent mechanism for cancer cells to silence expression of tumor
suppressor genes that can serve as the second “hit” to germ-line or somatic mutation[159,160].

3.1.3 DNA methylation in colorectal cancer
As described above, DNA methylation in CRC adopts a general pattern of genome-wide hypomethylation and hypermethylation of CpG islands (CGIs) in gene promoter regions. The consequences of CGI methylation in promoters are usually repression of gene transcription. This was initially thought to be due to steric inhibition of transcription factor binding, however more recent evidence suggests that the silenced state likely precedes DNA methylation as chromatin modifications inducing a condensed, and therefore repressed state, are present before DNA methylation[154]. Thus, it seems that the more likely role of CGI methylation in gene expression is long-term stabilization of the silenced state.

Regardless of the mechanism, gene silencing due to DNA methylation is a key event in CRC pathogenesis, serving as another means of gene inactivation. Targets frequently silenced in CRC by promoter hypermethylation include tumor suppressors, DNA repair genes, and cell adhesion proteins. Similar to the sequential accumulation of somatic mutation during the transformation of the normal colon epithelium into an invasive cancer, certain genes are often silenced by methylation at specific steps in the sequence. For example, aberrant WNT signaling, considered the initiating event in CRC, is usually due to somatic mutation and LOH of the APC gene[6]. However, methylation induced silencing of APC can act as the surrogate to somatic mutation and aberrant WNT signaling can be further exacerbated by methylation of WNT antagonist genes SFRP1.
and SFRP2. Promoter hypermethylation of p16 is also common in CRC tumors with activating KRAS and BRAF mutations, allowing cancer cells to circumvent oncogene-induced senescence by p16 activation\[161,162].

While all CRC tumors exhibit some degree of aberrant gene methylation, some have a higher frequency of recurrently methylated genes, a subgroup that has been called CpG Island methylator phenotype (CIMP)\[163]. About 20% of CRCs are CIMP tumors and associate with a unique molecular and histologic phenotype. Almost all CIMP tumors arise from the right-side of the colon in sessile serrated adenomas as opposed to traditional tubular or villous adenomas and occur more frequently in female patients\[164]. Further, these tumors exhibit a high frequency of concomitant BRAF mutation and methylation of the DNA mismatch repair gene MSH1\[165]. Loss of MLH1 expression gives rise to sporadic microsatellite instability (MSI), the same genomic defect seen in hereditary non-polyposis coli (HNPCC, aka Lynch syndrome), where germ-line mutations in MLH1 or MSH2 underlie MSI development. The prognosis for CIMP cancers is generally more favorable although they may differ based on further molecular sub-classification\[166].

3.1.4 Methylated tumor DNA as a Biomarker

Generally speaking, biomarkers are any biological molecule that signifies an abnormal process or disease condition\[167]. This could include proteins or nucleic acids in body fluids that could be used for disease monitoring or gene expression profiling of tumors to predict treatment response and clinical management. Typically we think of biomarker assays as an objective measure to differentiate affected patients from persons without
disease. It is in this latter context that we refer for using methylated DNA as a biomarker for cancer detection.

Currently, there are few minimally invasive assays used for detection and clinical monitoring of cancer recurrence in CRC patients. Screening for CRC is performed by colonoscopy, which carries a high cost and poor patient adherence to current screening guidelines[168]. The only regularly used blood-based biomarker in CRC is carcinoembryonic antigen (CEA) for monitoring disease recurrence. Assaying for CEA has limited screening utility because of low sensitivity and specificity. Further, CEA can be elevated during bouts of inflammation, such as in patients with inflammatory bowel conditions, and is prone to false-positive results[169]. As early detection is a major factor in successful cancer treatment, the paucity of minimally invasive biomarkers for CRC detection has made this a critical area of investigation.

The rationale for using aberrantly methylated DNA as a cancer biomarker spurs from the fact that methylation events are frequent and occur early on in CRC pathogenesis, that these changes are stable and, importantly, they can be detected in biological fluids[170]. As discussed previously, highly recurrent mutations in CRC occur only in a few genes, and even these mutation events do not approach the >90% frequency of hypermethylation events described at many CGI loci[171]. Furthermore, DNA methylation changes occur early on in the course of CRC at the adenoma stage, presenting an opportunity for early detection[172]. Lastly, tumor DNA can be detected in a variety of biological fluids, such as blood, urine, saliva, or stool, depending on the anatomical location of the tumor[173]. The ability to detect methylated tumor DNA in various biological samples makes this an attractive candidate biomarker for the clinic.
Work from our laboratory pioneered some of the earliest studies providing proof-of-concept on methylated biomarkers in CRC by demonstrating detection of two methylated genes in patient samples: MLH1 in plasma and methylated vimentin in stool of patients with CRC[174,175]. Since these reports, there have been numerous studies of other candidate methylated biomarkers in both plasma and stool, but many of these have been plagued by inconsistent results in duplicate studies or overall inadequate sensitivity and/or specificity for clinical use[176]. This is reflected by the fact that no plasma based detection methods are in clinical use and only one FDA approved test is currently on the market for CRC screening in stool DNA samples[177]. Though stool-based assays are non-invasive, they do not offer the convenience of collection and testing during a routine clinic visit and are still met with issues of patient non-compliance. Thus, for pragmatic purposes, discovery of a blood based biomarker remains as the best solution for non-invasive testing in CRC. We therefore sought to perform a genome-wide screen of the CRC methylome using reduced representation bisulfite sequencing to identify novel methylated loci in CRC tumors for development of a blood-based, methylated biomarker assay.

3.1.5 Reduced Representation Bisulfite Sequencing: a tool for examining genome-wide DNA methylation

The CpG distribution of the genome is highly uneven, consisting of short, dense stretches of CpGs dotting large swaths of CpG sparse regions. These CpG densities, mainly CpG islands and their surrounding CpG island shores, makeup less than ~2% of the genomic real estate. In seeking candidates for a methylated biomarker assay, one would ideally
sequence the CpG rich regions, such as the CpG islands and CpG island shores, as these regions are the most frequent targets of hypermethylation in tumors and will also be the most informative. Similar to the targeted enrichment approach whole exome sequencing discussed in the previous chapter, one can enrich for the CpG dense areas in the genome before sequencing to improve the efficiency. Reduced representation bisulfite (RRBS) sequencing is one such approach well suited for the discovery of differentially methylated loci[178].

The workflow for RRBS sample preparation is outlined in Figure 3.1. The key enrichment step is the enzymatic digestion of genomic DNA with MSPI and size selection of the MSPI digested fragments. The MSPI enzyme cleaves at CCGG (C\_CGG) sites and can digest methylated or unmethylated sequences alike. Two points to highlight are that the CCGG recognition sequence will occur more frequently in CG dense regions, resulting in multiple small fragments after MSPI digestion and every MSPI fragment will contain at least two terminal CpGs. Size selection of fragments in the range of 40-220bp gives representative coverage of most CpG islands and other CpG dense regions. Following size selection, bisulfite treatment will convert unmethylated cytosine to uracil while methylated cytosine will be protected from conversion. In this manner, one can decipher the methylation pattern of the original allele. Read mapping is conducted by aligning sequenced reads to a bisulfite converted copy of the genome and percent methylation at a given position is calculated by dividing the number of methylated calls by the total coverage of the position.

Overall, RRBS only uses an input of ~1% of total genomic DNA to provide a representative sampling of ~10% of the genome[179]. The method offers advantages of
high coverage of CpG dense regions at single base pair resolution while only needing around 10-20 million reads per sample for adequate read depth, providing affordability for large comparison studies.

3.2 Approach

We adapted RRBS to identify novel methylated loci in CRC for biomarker discovery. Using a discovery cohort of CRC tumors and matched normal colon we identified 100 candidate regions for further characterization. Validation of our candidates was carried out using a targeted resequencing approach, where individual PCR amplification of the 100 candidate regions was performed in an expanded discovery cohort and followed by deep sequencing. Each of the amplicons was assessed for sensitivity/specificity to discriminate tumor from normal. Our top performing candidates were then further scrutinized to identify windows: the best 70 bp regions within each amplicon containing the highest number of CpGs able to discriminate tumor from normal. The top windows were selected and PCR amplified in a second validation set to assess performance. These windows will be next amplified in a cohort of plasma samples to determine their potential as methylated plasma biomarkers for CRC detection and monitoring.

3.3 Materials and Methods

Sample Collection and DNA isolation
The tumor sample accrual protocol entitled, “CWRU 7296: Colon Epithelial Tissue Bank”, was approved by the University Hospitals Case Medical Center Institutional Review Board for Human Investigation with the assigned UH IRB number 03-94-105.
Under this protocol, discarded tissue was obtained through written informed consent from patients for research use. Tumor specimens were obtained from a frozen archive that consisted of sporadic colorectal cancers without reported family history accrued under the above protocol. Clinical data was obtained and assembled through individual pathology case reports for each tumor. Microscopic review of tumor morphology for selected samples was performed by an anatomic pathologist (J.W). Only specimens containing greater than 50% of tumor material were selected for analysis. Genomic DNA extraction from tumor samples was performed using either a standard guanidine thiocyanate protocol[58] or the DNeasy Blood and Tissue Kit (QIAGEN). Testing for MSI status at genomic loci BAT26 and BAT40 was performed as previously described[59].

Plasma samples were collected under an internal review board approved protocol at University Hospitals Case Medical Center. Samples were obtained and processed as previously described[180]. Briefly, 18 mls of pre-operatively collected blood was loaded into a standard blood collection tube containing EDTA, immediately chilled to 8°C, and transported for processing within 1 hr of collection. The blood cells were pelleted and plasma was transferred to 1.5 ml tubes, immediately frozen, and stored at -80°C. Plasma samples were thawed at 25°C for 5 min, centrifuged to remove debris, and transferred to a new tube. Total DNA was purified from the plasma using the QIAamp minElute Virus Kit (QIAGEN, 57714) as per the manufacturer’s instructions.

Reduced Representation Bisulfite Sequencing
Reduced representation bisulfite sequencing was performed as described by Gu et al. [181] Briefly, genomic DNA isolated from either frozen colon tumor tissue or normal colon mucosa was MSP1 digested, end repaired, and poly-A tailed. Adapter ligation was carried out using the NEXTFlex Rapid illumina DNA-Seq library prep kit (Bioo Scientific, cat# 5144-02) and using NEXTFlex-96 DNA barcoded adapters (Bioo Scientific, cat# 514105) to allow for multiplexing of samples on a single sequencing lane. Following adapter ligation, the library was size fractionated on an agarose gel and DNA fragments between 170-350 bps (corresponding to the adapter ligated 40-220 bp MSP1 digested fragments) were isolated and bisulfite converted using the EpiTect Bisulfite Conversion Kit (QIAGEN, cat# 59104). Adapter ligated and converted libraries were then PCR amplified and purified as outlined in the NEXTFlex Rapid Illumina DNA-Seq library prep kit and library quality assessed using an Agilent 2100 Bioanalyzer.

Sequencing was performed on an Illumina HiSeq 2500 using paired-end, 100bp reads. Individual BAM files for each sample were generated for alignment. DNA sequencing reads from each RRBS experiment were aligned to bisululfite converted and unconverted versions of the human reference genome (hg18) using Bowtie2, and percent methylation for each CpG was calculated by dividing the number of methylated Cs by the total coverage of that base. These analyses were facilitated by the Bismark software [PMID: 21493656], which was specifically developed for processing RRBS data. The overall pipeline converts raw RRBS fastq files to tables of read depth and percent methylation at each individual CpG site for each patient sample.
Filtering of Candidate Differentially Methylated Regions

The RRBS discovery set consisted of 25 unique normal samples and 41 unique tumors and yielded 3,091,193 CpGs in MspI fragments between 40 bp and 220 bp for analysis. For each normal sample, we only considered CpGs with read depth of at least 20 (the rationale for this is that, since tumor must differ substantially from all normals, stochasticity would be more likely to make a single normal sample appear to have very different methylation levels than it does, thereby omitting truly hyper- or hypo-sites in tumor). In each normal sample, a CpG (with depth >19) was classified as unmethylated (< 5%), hemi-methylated (40-60%), or methylated (>90%). If it had sufficient depth and was not in these categories, it was classified as intermediary and if depth was < 20, it was classified as “insufficient depth”.

A CpG was classified across normals as consensus as either unmethylated, hemi-methylated, or methylated if all samples were of sufficient depth (>19) and there were at least four of them, otherwise they were labeled as no consensus and not considered for further analysis. To identify differentially methylated regions of interest, a CpG was classified for each tumor sample as UpFromUn (Un_Up) if there was a normal unmethylated consensus and all consensus normals had methylation levels at least 20% points below that of the tumor; UpFromHemi (Hemi_Up) if there was a normal hemi-methylated consensus and all consensus normals had methylation levels at least 20% points below that of the tumor; DownFromMeth (Methyl_Down) if there was a normal methylated consensus and all consensus normals had methylation levels at least 20% points above that of the tumor; DownFromHemi (Hemi_Down) if there is a normal hemi-methylated consensus and all consensus normals had methylation levels at least 20%
points above that of the tumor; no change if normals had no consensus or tumor did not change at least 20% points from all (informative) normal; or lastly labeled as uninformative if falling outside of any of these categories. A minimum tumor read depth of >10 was required for classification of any given CpG. This classification process yielded a total of 44,681 individual CpGs differing in at least one tumor with the following breakdown by category: 8717 Un_Up CpGs, 38,672 Methyl_Down CpGs, 781 Hemi_Down CpGs, and 398 Hemi_Up CpGs.

To further filter out CpG regions for validation we required that for any CpG category that at least 80% of tumors with sufficient read depth differ from the normal consensus CpG state. We also sought to identify loci with two or more informative CpGs within a 200 bp region, what we defined as a “patch”. Thus, for any given category, we selected those CpGs occurring in a patch where at least one CpG within this patch differed in greater than 80% of tumors for validation by targeted resequencing. Using these criteria, we identified 100 candidate patches for further validation with the following breakdown by category: 50 Un_Up patches, 35 Methyl_Down patches, 13 Hemi_Down patches, and 2 Hemi_Up patches.

**Loci Validation and Targeted Resequencing**

For targeted resequencing of candidate patches, we designed methylation independent primers against *in silico* bisulfite converted sequence for each of the 100 patches of interest. Primer design was optimized for a maximum amplicon size of approximately 220 bp to allow for greatest sequencing coverage and maximal CpG inclusion within
each patch. The list of primer sequences and their corresponding genomic loci are listed in Appendix Table 1.

The validation sample set consisted of 40 tumors from the discovery set and their matched normal colon tissue as well as 8 additional tumor samples with their derivative colon cancer cell lines, totaling 96 samples for validation. Bisulfite conversion of genomic DNA was carried out as above using the EpiTect Bisulfite Conversion Kit 48 (QIAGEN, cat# 59104). PCR amplification of each locus was performed using Platinum® Taq DNA Polymerase (Life Technologies, cat# 10966-083). Each 50μl reaction consisted of a final concentration of 2.5 U DNA polymerase, 2 mM MgCl$_2$, 0.2 mM dNTPs, 0.2 mM each of forward and reverse primer, and 50 ng of input bisulfite converted DNA. Cycling conditions for PCRs followed a touchdown protocol consisting of an initial denaturing step at 95°C for 5 min, then 3 cycle increments beginning with 95°C for 45 s, 64°C for 45 s, 72°C for 30 s where annealing temperature was reduced by 3°C at the completion of each increment until the annealing temperature reached 52°C in which the last amplification cycle consisted of 25 cycles of 95°C for 45 s, 52°C for 45 s, and 72°C for 30 s with a final elongation step of 72°C for 10 min.

For each sample, all PCR products were visualized on an agarose gel to verify successful amplification. Following visualization, PCR products were combined and column purified using the NucleoSpin Gel and PCR Clean-up kit (Clontech, cat# 740609). Sample concentration of the purified products was obtained using the Qubit dsDNA broad range assay kit (Life Technologies, cat# Q32853). Library preparation was performed as described in the NEXTFlex Rapid illumina DNA-Seq library prep kit (Bioo Scientific, cat# 5144-02) user manual using 1 ug of input DNA for each sample and the
NEXTFlex-96 DNA barcoded adapters (Bioo Scientific, cat# 514105). Samples were then sequenced on an Illumina HiSeq 2500 using paired end, 100 bp reads. BAM files for each sample were aligned and analyzed as above.

**Selection of Best Candidate Windows and Window Validation**

For each CpG in the regions of interest, we defined the "background" to be the second largest methylation proportion among normal samples (that is, we allow one outlier). Also, for each CpG in the regions of interest, we define the "signal" to be the average of the median methylation proportion for tumor samples and the median methylation proportion for cell line samples. For each 70 bp window, we consider only CpGs with background less than 0.1; the "score" for the window is the sum of the signals for those CpGs with background less than 0.1. (This score is equivalent to the average signal among CpGs with good background times the number of CpGs with good background). We selected the 5 best windows as the 5 windows with highest scores. We then tested each window for sensitivity and specificity and chose the window from each group with the highest sensitivity given maximal specificity.

The validation cohort consisted of an independent set of 96 tumors with their matched normal colon tissue. Samples were bisulfite converted and windows PCR amplified as above using the specified primers in Appendix Table 2 Following amplification the workflow for purification, library prep, sequencing, and alignment was as described for the loci validation above.
3.4 Results

3.4.1 Reduced Representation Bisulfite Sequencing and Variant Filtering to Identify Candidate Methylated Loci

We performed RRBS on set of 66 samples comprised of 41 unique colon tumors and 25 matched normal colon epithelium to identify differentially methylated loci for novel biomarker discovery. From this experiment, over 3 million CpGs were identified for analysis. We filtered candidate CpGs following the schematic outlined in Figure 3.2 (criteria defined in Materials and Methods). We first excluded CpGs with insufficient read depth (<20 in normal samples) for analysis then categorized remaining CpGs based on their consensus methylation state in normal colon tissue. We defined three possible bins for categorization: unmethylated (<5%), hemi-methylated (40-60%), and methylated (>90%) and excluded all CpGs falling outside of these bins. Given these three categories for normal methylation state, we defined four possible comparisons to the methylation change in colon cancer tumors as follows: Un_Up where the normal consensus CpG is unmethylated and becomes methylated in tumor; Methyl_Down where the normal consensus CpG is methylated and becomes unmethylated in tumor; Hemi_Up where the normal consensus CpG is hemi-methylated and becomes further methylated in tumor; and Hemi_Down where the normal consensus CpG is hemi-methylated and becomes demethylated in tumor.

In each category, we only considered CpG sites differing in tumors by more than 20%. This yielded over 48,000 CpGs from all four categories differing from normal in at least one tumor. We further filtered CpGs for each comparison group by requiring that a
CpG differ from normal in at least 80% of tumors and that these CpGs occur within 200 bps of another informative CpG, an area that we defined as a “patch” (Figure 3.3). This final filtering step left us with 100 candidate methylated patches for validation: 50 Un_Up, 2 Hemi_Up, 13 Hemi_Down, and 35 Methyl_Down patches.

3.4.2 Patch Validation identifies 4 primary methylated candidate biomarkers

To validate the identified patches, we performed PCR amplification of bisulfite converted genomic DNA for targeted deep sequencing in an expanded discovery cohort consisting of 96 total samples: 40 colon tumors and their matched normal colon sample and 8 colon cancer cell lines and their parental tumors. We chose to perform resequencing on an expanded discovery set as opposed to an independent tumor cohort as the initial RRBS sequencing had incomplete coverage in many of the samples. PCR assays for targeted resequencing were designed to include the patch and as many CpGs within a ~220 bp amplicon (Figure 3.3).

Sensitivity and specificity of each amplicon were analyzed using a methylation-per-read analysis that can distinguish epialleles by the methylation pattern (Figure 3.4). To graph sensitivity and specificity of each amplicon, we required that the number of CpGs counted towards methylation be present in at least 1% of reads. We sought amplicons with highest sensitivity given maximal specificity. From this analysis, we found four priority candidates for distinguishing between normal and tumor: Un_Up_106, Un_Up_146, Un_Up_207 and Un_Up_307 (Figure 3.5). Sensitivity and specificity for each amplicon were as follows (sens/spec): Un_Up_106 (79%/100%); Un_Up_146 (92%/95%); Un_Up_207 (69%/100%); Un_Up_307: (80%/100%).
3.4.3 Window Identification in top candidate biomarkers

Circulating DNA in cancer patients is highly degraded and PCR amplification of DNA fragments is most efficient for smaller amplicons (~100 bp)[182]. Given that our amplicons for targeted resequencing averaged around 200bp, we performed a bioinformatic scan of our top candidate amplicons to identify the most informative 70bp interval, an area we defined as a “window”, for further validation and future use in plasma (Figure 3.3). For our candidate amplicons we analyzed the five best windows for sensitivity/specificity and identified the top window for each amplicon in our expanded discovery set (sens/spec): Un_Up_106: (85%/100%); Un_Up_146: (95%/95%); Un_Up_207 (81%/100%); Un_Up_307: (87%/100%) (Figure 3.6).

3.4.4 Validation of Candidate Windows in an independent tumor cohort

We next performed an independent validation of our candidate windows in cohort of 96 CRC tumors and their matched normals (Figure 3.7). The performance of Un_Up_146 was comparable to the discovery cohort with an overall sensitivity/specificity of 82%/92% in the validation set. The sensitivity/specificity for Un_Up_207 was 40%/90% and Un_Up_307 3%/99%, thus they exhibited varied performance compared to the discovery cohort. Surprisingly, originally one of the best candidates, Un_Up_106 failed to validate. Thus, we will exclude Un_Up_106 from further analysis when we move into testing candidate windows in plasma of CRC patients.
3.5 Conclusions

In conclusion, we describe using RRBS to identify novel methylated loci in CRC for biomarker discovery. Using stringent filtering criteria, we were able to narrow down the field of candidate differentially methylated loci to 100 patches for validation. Targeted next generation resequencing of these candidate patches identified 4 primary patch candidates with high specificity and sensitivity for discerning tumor from normal. We identified the most informative 70 bp regions, the windows within each patch, for further validation and future use in plasma samples. Performance of the 4 windows in an independent tumor cohort was variable with only 1 window exhibiting comparable performance to the discovery set, two windows showing reduced performance, and one completely failing to validate. We will next test our remaining 3 candidate windows in a retrospective cohort of plasma samples to determine their performance as a blood based biomarker for CRC detection.

3.6 Discussion and Future Directions

The preliminary results from our work seeking to identify blood-based methylated biomarkers for CRC are encouraging. We have identified at least one reliable marker, Un_Up_146, exhibiting an overall sensitivity/specificity of 85%/93% in the two independent tumor cohorts combined. The performance of Un_Up_146 rivals that of other promising single methylated biomarkers in CRC[183]. Although our other candidate biomarkers were highly variable in the validation set with Un_Up_106 failing to validate, it still may be possible to use Un_Up_207 and/or Un_Up_307 in a panel
given their high specificity. Validation of candidates is a recurring problem in biomarker research and warrants further discussion.

The decreased performance of the other markers in the validation cohort is possibly due to overfitting of the data when we selected our candidate patches and windows. Overfitting is a statistical concept of finding a discriminatory pattern by chance[184]. In our case, we have a methylation pattern in a given locus, Un_Up_106 for example, that can distinguish between tumor and normal accurately and with high sensitivity of detection in our discovery set. However, when we expand this out to the validation set, it no longer performs as well. The reason being is that the discovery set is just that: a sample of the larger population. So in selecting our best candidate windows we chose those CpGs most informative for discriminating between tumor and normal in our training set (i.e. the discovery set) and by doing so, may have biased ourselves to selecting noise in this sample instead of a real biological difference in the population at large. Although we established criteria to reduce the possibility of a stochastic signal, such as requiring greater read depth in our normal comparison groups, having all normal samples in agreement and so forth, in large data sets where one is searching for discriminatory patterns there is always a risk of finding differences by chance alone. Improvements to the initial discovery experiment may have included a larger set of normal tissue for our comparison group or requiring greater read depth within these samples to decrease the chance of including noise in the comparison group.

Another possible confounding factor in any biomarker research is disease heterogeneity[185]. All solid tumors exhibit intratumoral heterogeneity, where the cellular constituents are a mix of stromal, inflammatory, and epithelial cells. In epithelial
tumors the non-epithelial cell populations are potential contributors to a noisy signal[186]. This can also be potential problem in the normal comparison group. Depending on the site of biopsy, adjacent normal tissue can be inflamed or subject to field cancerization, a phenomenon where ostensibly non-cancerous tissue surrounding tumors harbors particular epigenetic and genetic changes associated with neoplastic progression[187]. These potential sources of aberrantly methylated DNA in the normal comparison group might lead one to mistakenly exclude a potential candidate methylated region because of misperceived background noise in normal samples.

Heterogeneous methylation is another potential problem when assessing potential biomarker candidates. During aberrant CGI methylation in cancer, heterogeneous methylation is frequently observed[188]. As was alluded to earlier in the per-read analysis (Figure 3.4), the complement of methylated DNA in a biological sample consists of a mix of alleles that differ by their pattern of methylation, sometimes referred to as an epiallele. The biology of heterogeneous methylation is not well understood, but may result from stochastic or gradual changes during replication or flux in the DNA methylation patterns in the mitotic progeny of a cell[189]. At any rate, the presence of many epialleles from heterogeneous methylation illustrates the importance of identifying the most stable and informative fraction of methylated CpGs within a CGI for cancer detection. This was the impetus behind identifying patches with a large number of informative CpGs and subsequently finding the best windows for plasma detection.

Despite these issues, there is still great promise for using methylated DNA as a biomarker for cancer detection. Hypermethylation at CGIs is a highly frequent event in CRC and densely uniform within CGIs even though there are distinct methylation
patterns that can arise. However, this is most likely attributable to gains in methylation over time as demethylation is thought to be increasingly rare[188]. Therefore it seems reasonable to assume an initial methylation event occurs in a tumor-initiating cell that is propagated with each cell division and expanded over time. Further, these methylation changes are stable and maintained in metastatic disease, indicating they can be used to detect disease progression and recurrence[190].

Improvements in detection through digital methods such as next generation sequencing offer the advantages of per-read analysis for epiallele quantification and higher sensitivity for detection in biological fluids. Indeed, detection of as little as 1 molecule of methylated DNA in a background of 10000 unmethylated molecules (0.01%) can be achieved with deep sequencing[180]. Increased sensitivity is vital for early detection as methylated tumor DNA in plasma is relatively low in earlier disease stages but increases with progression[182]. Further, the ability to quantify individual epialleles allows for detection of densely methylated tumor in a background of sparsely methylated DNA. This will be highly important when testing our biomarker performance in the retrospective plasma cohort. The results of the plasma experiment will be the most telling as to the validity of our filtering approach and the overall success of the method.
3.7 Figures and Tables

**Figure 3.1. Workflow for RRBS library preparation.** The enrichment steps for CpG dense regions are MSPI digestion and size selection of fragments. Bisulfite treatment converts unmethylated cytosines to uracil whereas methylated cytosines will be protected, allowing the methylation pattern to be deciphered.
Figure 3.2. Schematic of Work Flow for Filter Candidate CpGs. The diagram illustrates the sequential classification and elimination of CpG candidates for further validation. From over 3 million CpGs returned for analysis, we identified a final set of 100 CpG patches for targeted resequencing. Filtering criteria are described in detail in Materials and Methods.
Figure 3.3. Diagram Illustrating Patches, PCR Amplicons, and Windows. Patches were defined as more than two CpGs within 200 bps where one of the CpGs was a candidate identified in filtering. The patch above shows three CpGs within a 120bp. The PCR amplicons for validation were designed to include the CpGs of the patch (red circles) and any other possible CpGs not included during the CpG filtering (yellow circles). These CpGs were not originally identified in the patch due to either a lack of coverage or they did not meet the defined filtering criteria. Validated amplicons were then scanned for windows, 70 bp regions with the highest proportion of informative CpGs in discriminating tumor from normal. The best window was the one having the most informative number of CpGs that could discriminate tumor from normal.
Figure 3.4. Schematic of methylation per-read versus methylation per CpG analysis.

The top panel shows methylation per CpG analysis can effectively discern between tumor and normal in a tumor biopsy sample. When analyzed in plasma however, due to the tumor DNA being largely diluted in a background of normal, this analysis is less sensitive (Bottom). Per-read analysis allows for identification of tumor DNA by requiring a minimum number of CpGs be methylated on a given read to be called positive. This approach is effective as tumor methylation reliably shows clusters of CpGs to be methylated as opposed to stochastic background methylation seen in normal samples.
Figure 3.5. Sensitivity and Specificity Graphs of Top 4 Candidate Patches from Targeted Resequencing in Expanded Discovery Set. Sensitivity/Specificity for amplicons were as follows: Un_Up_106: 79%/100%; Un_Up_146: 92%/95%; Un_Up_207 69%/100%; Un_Up_307: 80%/100%
Figure 3.6. Sensitivity and Specificity Graphs of Top 4 Candidate Windows from Targeted Resequencing in Expanded Discovery Set. Sensitivity/Specificity for amplicons were as follows: Un_Up_106: 85%/100%; Un_Up_146: 95%/95%; Un_Up_207 81%/100%; Un_Up_307: 87%/100%.
CpGs counted towards methylation

Un_Up_106 Window

Sensitivity
Specificity

Un_Up_146 Window

Sensitivity
Specificity

Un_Up_207 Window

Sensitivity
Specificity

Un_Up_307 Window

Sensitivity
Specificity
Figure 3.7. Sensitivity and Specificity Graphs of Top Window Performance in Validation Set. Un_Up_146 had a similar performance in the validation set with a sensitivity of 82% at a maximal specificity of 92%. Un_Up_207 and Un_Up_307 showed variable performance with sensitivity/specificity of 40%/90% for Un_Up_207 and 3%/99% for Un_Up_307. Un_Up_106 failed to validate.
## Appendix Table 1. Primer Sequences and Coordinates for Patch Amplification and Targeted Resequencing.

<table>
<thead>
<tr>
<th>For PrimerID</th>
<th>For Primer Seq</th>
<th>Rev PrimerID</th>
<th>Rev Primer Seq</th>
<th>Chr</th>
<th>For Primer Start</th>
<th>Rev Primer Start</th>
<th>Amplicon Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hemi_Do wn 2 F</td>
<td>GAGATTATGAGTTTAAYTGAGGGAAGTGA</td>
<td>Hemi_Do wn 2 R</td>
<td>TCCCCAAAATAAAATTCCAAATAGACAATTC</td>
<td>ch  r1</td>
<td>14922558</td>
<td>14922669</td>
<td>111</td>
</tr>
<tr>
<td>Hemi_Do wn 33 F</td>
<td>GGGATGATAGTTATTGAGGGTTTAGGAG</td>
<td>Hemi_Do wn 33 R</td>
<td>CAAAATATCTCTATCTAACCCT</td>
<td>ch  r7</td>
<td>56935807</td>
<td>56935959</td>
<td>152</td>
</tr>
<tr>
<td>Hemi_Do wn 34 F</td>
<td>GGTAGATTATGAGTTAGAGGTTAGGAG</td>
<td>Hemi_Do wn 34 R</td>
<td>CTCAACTCTACAACTTACCC</td>
<td>ch  r7</td>
<td>67093025</td>
<td>67093203</td>
<td>178</td>
</tr>
<tr>
<td>Hemi_Do wn 39 F</td>
<td>GAGATGATGAGTGGTTAGGAGATGGTTAGGAG</td>
<td>Hemi_Do wn 39 R</td>
<td>TCATAACTCTACAACTTACCC</td>
<td>ch  r8</td>
<td>50744876</td>
<td>50745047</td>
<td>146</td>
</tr>
<tr>
<td>Hemi_Do wn 40 F</td>
<td>GAGGATGATGAGTTAGGAGTTAGGAG</td>
<td>Hemi_Do wn 40 R</td>
<td>TAAAACCTCAAAATACTTACC</td>
<td>ch  r8</td>
<td>14255872</td>
<td>14255887</td>
<td>2</td>
</tr>
<tr>
<td>Hemi_Do wn 43 F</td>
<td>TAAAGATGATATGAGTTAGGAGTGGTTAGGAG</td>
<td>Hemi_Do wn 43 R</td>
<td>AAACCTCAAATTACCAACTCT</td>
<td>ch  r9</td>
<td>11490886</td>
<td>11491067</td>
<td>181</td>
</tr>
<tr>
<td>Hemi_Do wn 56 F</td>
<td>GGGTATGATGAGTTAGGAGTTAGGAG</td>
<td>Hemi_Do wn 56 R</td>
<td>TCCCCCTTCAACCTCACTT</td>
<td>ch  r2</td>
<td>34385563</td>
<td>34385722</td>
<td>159</td>
</tr>
<tr>
<td>Hemi_Do wn 57 F</td>
<td>AGGGATGATAGTTATTGAGTTAGGAG</td>
<td>Hemi_Do wn 57 R</td>
<td>CAAAATATCTCTACCACTTACCT</td>
<td>ch  r2</td>
<td>34387001</td>
<td>34387207</td>
<td>206</td>
</tr>
<tr>
<td>Hemi_Do wn 61 F</td>
<td>GGTAGTTAGATGGTTAGGAGTTAGGAG</td>
<td>Hemi_Do wn 61 R</td>
<td>CTCAACTCTACAACTTACCC</td>
<td>ch  r2</td>
<td>13148614</td>
<td>13148632</td>
<td>2</td>
</tr>
<tr>
<td>Hemi_Do wn 79 F</td>
<td>GTGGGATTAGGGAAGTGAAGTGA</td>
<td>Hemi_Do wn 79 R</td>
<td>CRACCACAAAAATTACCACTTACCT</td>
<td>ch  r9</td>
<td>1241752</td>
<td>1241892</td>
<td>140</td>
</tr>
<tr>
<td>Hemi_Do wn 82 F</td>
<td>GGGTATGATGAGTTATTGAGTTAGGAG</td>
<td>Hemi_Do wn 82 R</td>
<td>CAAAATATCTCTACCACTTACCC</td>
<td>ch  r9</td>
<td>38808380</td>
<td>38808528</td>
<td>148</td>
</tr>
<tr>
<td>Hemi_Do wn 91 F</td>
<td>GGTAGTTAGATGGTTAGGAGTTAGGAG</td>
<td>Hemi_Do wn 91 R</td>
<td>CCRACAACACCTTACCTTACCT</td>
<td>ch  r2</td>
<td>46676499</td>
<td>46676668</td>
<td>169</td>
</tr>
<tr>
<td>Hemi_Do wn 95 F</td>
<td>GGGATGAAATATGAGGGAAGTGA</td>
<td>Hemi_Do wn 95 R</td>
<td>AACCCAAAAACCTCAACCA</td>
<td>ch  r2</td>
<td>46038367</td>
<td>46038541</td>
<td>174</td>
</tr>
<tr>
<td>Hemi_Up 12 F</td>
<td>AGGTTAGTTAGGAGTTAGGAG</td>
<td>Hemi_Up 12 R</td>
<td>CAACCAACTCTCTAACCAA</td>
<td>ch  r6</td>
<td>601234</td>
<td>601437</td>
<td>203</td>
</tr>
<tr>
<td>Hemi_Up 17 F</td>
<td>GTGGTATGATGAGTTATTGAGTTAGGAG</td>
<td>Hemi_Up 17 R</td>
<td>CTCAACTCTACAACTTACCTTACCT</td>
<td>ch  r2</td>
<td>61669277</td>
<td>61669413</td>
<td>136</td>
</tr>
<tr>
<td>Methyl_Do wn 102 F</td>
<td>TTGTTATGATGAGTTATTGAGTTAGGAG</td>
<td>Methyl_Do wn 102 R</td>
<td>TAAAACCTCAAAACACTACCAAA</td>
<td>ch  r3</td>
<td>19696933</td>
<td>19696950</td>
<td>3</td>
</tr>
<tr>
<td>Methyl_Do wn 105 F</td>
<td>GAGAAATAGAGTTAGGAGGTTAGGAG</td>
<td>Methyl_Do wn 105 R</td>
<td>CAACCAACACTTACCACCAAA</td>
<td>ch  r4</td>
<td>1289479</td>
<td>1289596</td>
<td>220</td>
</tr>
<tr>
<td>Methyl_Do wn 112 F</td>
<td>GGGGAATGTTAAGTTATTGAGTTATTGAG</td>
<td>Methyl_Do wn 112 R</td>
<td>CCACTATCTCTATACCCCAACCAA</td>
<td>ch  r4</td>
<td>56402662</td>
<td>56402811</td>
<td>149</td>
</tr>
<tr>
<td>Methyl_Do wn 118 F</td>
<td>AAGGGTGGTATGAGTTATTGAGTTATTGAG</td>
<td>Methyl_Do wn 118 R</td>
<td>ATCTATCTCTATACCCCAACCAA</td>
<td>ch  r5</td>
<td>11415572</td>
<td>11415594</td>
<td>220</td>
</tr>
<tr>
<td>Methyl_Do wn 12 F</td>
<td>GTGGTATGATGAGTTATTGAGTTATTGAG</td>
<td>Methyl_Do wn 12 R</td>
<td>CAACCAACACTTACCACCAAA</td>
<td>ch  r5</td>
<td>43246309</td>
<td>43246518</td>
<td>209</td>
</tr>
<tr>
<td>Methyl_Do wn 132 F</td>
<td>GAGGAGGAAAATAGGGTTATGAGTTATTGAG</td>
<td>Methyl_Do wn 132 R</td>
<td>AAATCTACCTTATACCCCAACCAA</td>
<td>ch  r5</td>
<td>39483903</td>
<td>39483924</td>
<td>121</td>
</tr>
<tr>
<td>Methyl_Do wn 141 F</td>
<td>GAGAATGATGAGTTATTGAGTTATTGAG</td>
<td>Methyl_Do wn 141 R</td>
<td>CCACTATCTCTATACCCCAACCAA</td>
<td>ch  r5</td>
<td>13018133</td>
<td>13018154</td>
<td>208</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTTGGGGTTGTTTAGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TAACCTCRTCAACATGCTTATTCATAAAC</td>
<td>ch</td>
<td>r6</td>
<td>14732541</td>
<td>14732698</td>
</tr>
<tr>
<td>----------</td>
<td>-----------------------------</td>
<td>------------</td>
<td>---------------------------------</td>
<td>-----</td>
<td>----</td>
<td>-----------</td>
<td>-----------</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGGTTTATTGAGTTTAGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TTACAAAATACATCTACACTACCTAACAC</td>
<td>ch</td>
<td>r6</td>
<td>97678116</td>
<td>97678300</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>CAATCTCCACCTCCCTATTCTCAAC</td>
<td>ch</td>
<td>r6</td>
<td>1531921</td>
<td>15319359</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TACCRACTACCTCCTACACTCTA</td>
<td>ch</td>
<td>r7</td>
<td>17033429</td>
<td>17033440</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TACACTACCTCCTACACTCTA</td>
<td>ch</td>
<td>r7</td>
<td>90107151</td>
<td>90107287</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>CACAAAACCTACCTCTTATTCATAACAC</td>
<td>ch</td>
<td>r8</td>
<td>42674520</td>
<td>42674712</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TTACACTACCTCCTACACTCTA</td>
<td>ch</td>
<td>r8</td>
<td>11015262</td>
<td>11015273</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TAACAAAAAACACTCAAAAACTCAACTCC</td>
<td>ch</td>
<td>r8</td>
<td>11843232</td>
<td>11843252</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>CCCTCAAAATAACTCACACTCAACTCC</td>
<td>ch</td>
<td>r8</td>
<td>68029485</td>
<td>68029662</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TACACTACCTCCTACACTCTA</td>
<td>ch</td>
<td>r8</td>
<td>11508698</td>
<td>11508719</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TACCTACCTCCTACACTCTA</td>
<td>ch</td>
<td>r8</td>
<td>27730994</td>
<td>2773299</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>ACATAACACTCAACTCCACCACT</td>
<td>ch</td>
<td>r9</td>
<td>22361061</td>
<td>22361301</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TCTACTACCTCCTACACTCACTCCAACT</td>
<td>ch</td>
<td>r9</td>
<td>22360886</td>
<td>22361081</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TATCTACTACCTCCTACACTCACTCCAA</td>
<td>ch</td>
<td>r9</td>
<td>18858314</td>
<td>18858531</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>AAGAGTTGTTTTGGAGTATGAGGAT</td>
<td>ch</td>
<td>r9</td>
<td>34859916</td>
<td>34860601</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>AAGATTTTATAGGTTAGGTT</td>
<td>ch</td>
<td>r9</td>
<td>46584154</td>
<td>46584366</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>TTTTATGTTGATGTTAG</td>
<td>ch</td>
<td>r9</td>
<td>20085821</td>
<td>20085921</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>GCACAACTAACATACATCTAAACAC</td>
<td>ch</td>
<td>r9</td>
<td>53594533</td>
<td>53594703</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>ATCTACAATCTAACACTACACTCACTCA</td>
<td>ch</td>
<td>r9</td>
<td>41793208</td>
<td>41793408</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>CCAATTTCTACATTGCTATAATAATAAC</td>
<td>ch</td>
<td>r9</td>
<td>32530329</td>
<td>32530520</td>
</tr>
<tr>
<td>Methyl_D</td>
<td>GGTAGTTGAGTAGTGGTGGAGTAGT</td>
<td>Methyl_Do</td>
<td>CCAATTTCTACATTGCTATAATAATAAC</td>
<td>ch</td>
<td>r9</td>
<td>24190372</td>
<td>24190388</td>
</tr>
<tr>
<td>Methyl D own 48 F</td>
<td>AGATAGTGCGTGTCGAGTCGT</td>
<td>Methyl D own 48 R</td>
<td>CCTGCCAACCAAATATAATAATATTCTCCT</td>
<td>ch r2</td>
<td>21290753</td>
<td>21290984</td>
<td>231</td>
</tr>
<tr>
<td>Methyl D own 51 F</td>
<td>GGGGATTTTTGGAATTTTTTTGTGAA</td>
<td>Methyl D own 51 R</td>
<td>CTTAAACCTAAAATATATACCTCT</td>
<td>ch r2</td>
<td>43266152</td>
<td>43266275</td>
<td>123</td>
</tr>
<tr>
<td>Methyl D own 55 F</td>
<td>TTTGATTTTTGGGAGAATTAG</td>
<td>Methyl D own 55 R</td>
<td>CTACCTGAAGCTCCTACT</td>
<td>ch r2</td>
<td>61243997</td>
<td>61244152</td>
<td>155</td>
</tr>
<tr>
<td>Methyl D own 62 F</td>
<td>GTATTTTGGGAGGGAATTTTTTTTTTAG</td>
<td>Methyl D own 62 R</td>
<td>CCAAAACATTAACTATATACCTCT</td>
<td>ch r2</td>
<td>14366475</td>
<td>14366494</td>
<td>0</td>
</tr>
<tr>
<td>Methyl D own 85 F</td>
<td>AAAAGTGGTTGGGAAATTTTTTTTTTTTTTTTGTGGA</td>
<td>Methyl D own 85 R</td>
<td>TACRATCAAACCTCTACTACAACTCCTC</td>
<td>ch r2</td>
<td>31939860</td>
<td>31939872</td>
<td>112</td>
</tr>
<tr>
<td>Methyl D own 90 F</td>
<td>AGAGAGGGGtttTAATAATTTTTTGGGAGAATT</td>
<td>Methyl D own 90 R</td>
<td>TAACCTGCCTTTACCCACCTC</td>
<td>ch r3</td>
<td>10760766</td>
<td>10760776</td>
<td>9</td>
</tr>
<tr>
<td>Un_Up_10 0 F</td>
<td>TGGGGAGGGAGGGTGTGATTTG</td>
<td>Un_Up_10 0 R</td>
<td>AACTTCCTCCACCCACTCTCACAAC</td>
<td>ch r6</td>
<td>11833595</td>
<td>11833617</td>
<td>227</td>
</tr>
<tr>
<td>Un_Up_10 2 F</td>
<td>TACGGGGTTATATTGCCGAGG</td>
<td>Un_Up_10 2 R</td>
<td>ATCTCCCTGAACCTCAATGAATCGC</td>
<td>ch r7</td>
<td>15032720</td>
<td>15032721</td>
<td>6</td>
</tr>
<tr>
<td>Un_Up_10 4 F</td>
<td>GTGATTTTATGGTATGAGG</td>
<td>Un_Up_10 4 R</td>
<td>TTGACCTACCAATCTATCTATCTTAC</td>
<td>ch r7</td>
<td>15217109</td>
<td>15217119</td>
<td>6</td>
</tr>
<tr>
<td>Un_Up_10 6 F</td>
<td>GTTACGTGTTATGGGTAGG</td>
<td>Un_Up_10 6 R</td>
<td>TCTACCTCTTACCTCTCCTATTACCTATTAAACCCATC</td>
<td>ch r6</td>
<td>16375474</td>
<td>16375493</td>
<td>2</td>
</tr>
<tr>
<td>Un_Up_10 7 F</td>
<td>GGGGTTATAAAGGAGGAGTGT</td>
<td>Un_Up_10 7 R</td>
<td>AACCTACCTGCCACCCACCTACTTACATCTATCTTACCTATTACCTATTAAACCCATC</td>
<td>ch r6</td>
<td>16375649</td>
<td>16375663</td>
<td>2</td>
</tr>
<tr>
<td>Un_Up_11 2 F</td>
<td>GATTGAGGAATTGTGTTAGGAAG</td>
<td>Un_Up_11 2 R</td>
<td>CTCTACACACCTTCTCSTACCCCTAC</td>
<td>ch r7</td>
<td>35192933</td>
<td>35193092</td>
<td>159</td>
</tr>
<tr>
<td>Un_Up_11 8 F1</td>
<td>TTATTAGGTGTTGTTAGGAGAATTAAGGAAAT</td>
<td>Un_Up_11 8 R</td>
<td>CTACCTACAACTTCTACCCCTACCCCTAC</td>
<td>ch r7</td>
<td>88267249</td>
<td>88269086</td>
<td>159</td>
</tr>
<tr>
<td>Un_Up_11 8 F2</td>
<td>GTTAGTTAGTTATGTTGTAAGAATTAGGTTAAGTATGATGTAAGTATGATGTAAGTATGATGTAAGTATGATGTAAGTATGATGTAAGTATGATGTAAGTATGATGTAAGTATGATGTAAGTATGATG</td>
<td>Un_Up_11 8 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r7</td>
<td>88267583</td>
<td>88269778</td>
<td>195</td>
</tr>
<tr>
<td>Un_Up_12 8 G1</td>
<td>GTTTAGGTTTTATAGGATGAATATTAGGAAAT</td>
<td>Un_Up_12 8 R</td>
<td>TCACCCACACACCTACCCCTACCCCTAC</td>
<td>ch r7</td>
<td>14220492</td>
<td>14220510</td>
<td>9</td>
</tr>
<tr>
<td>Un_Up_12 8 G2</td>
<td>GATTTAGATTTAAGTTAGGATGAATATTAGGAAAT</td>
<td>Un_Up_12 8 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r7</td>
<td>15073839</td>
<td>15073855</td>
<td>1</td>
</tr>
<tr>
<td>Un_Up_13 8 G</td>
<td>GGGTGGATTTGTTAGGATGAATATTAGGAAAT</td>
<td>Un_Up_13 8 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r7</td>
<td>41873983</td>
<td>41874204</td>
<td>221</td>
</tr>
<tr>
<td>Un_Up_14 3 F</td>
<td>GGTACGTGTTATGGGTAGG</td>
<td>Un_Up_14 3 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r8</td>
<td>68037369</td>
<td>68037524</td>
<td>155</td>
</tr>
<tr>
<td>Un_Up_14 5 F</td>
<td>GGGTACGTGTTATGGGTAGG</td>
<td>Un_Up_14 5 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r8</td>
<td>97242641</td>
<td>97242781</td>
<td>140</td>
</tr>
<tr>
<td>Un_Up_14 6 F</td>
<td>GATTGATAGATTTATAGGATGAATATTAGGAAAT</td>
<td>Un_Up_14 6 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r8</td>
<td>9757569</td>
<td>97575857</td>
<td>166</td>
</tr>
<tr>
<td>Un_Up_15 9 F</td>
<td>GTTACGTGTTATGGGTAGG</td>
<td>Un_Up_15 9 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r8</td>
<td>14353029</td>
<td>14353043</td>
<td>2</td>
</tr>
<tr>
<td>Un_Up_16 5 F</td>
<td>GGGGGAACGAGGTAGTGGAAAGG</td>
<td>Un_Up_16 5 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r8</td>
<td>10107956</td>
<td>10107975</td>
<td>7</td>
</tr>
<tr>
<td>Un_Up_17 7 F</td>
<td>GGGTACATGAAGGAGTATTCCGAGT</td>
<td>Un_Up_17 7 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r1</td>
<td>627058</td>
<td>627264</td>
<td>206</td>
</tr>
<tr>
<td>Un_Up_19 0 F</td>
<td>GGGTACGTGTTATGGGTAGG</td>
<td>Un_Up_19 0 R</td>
<td>ATCTACCCACCTTCTACCCCTACCCCTAC</td>
<td>ch r1</td>
<td>12280629</td>
<td>12280645</td>
<td>4</td>
</tr>
<tr>
<td>Un_Up_20 2 F</td>
<td>GGGTACGTGTTATGGGTAGG</td>
<td>Un_Up_20 2 R</td>
<td>CTACCTACACACCTTCTACCCCTACCCCTAC</td>
<td>ch r1</td>
<td>63505375</td>
<td>63505529</td>
<td>154</td>
</tr>
<tr>
<td>Un_Up_20 4 F</td>
<td>GGGTACGTGTTATGGGTAGG</td>
<td>Un_Up_20 4 R</td>
<td>ATCTACCCACCTTCTACCCCTACCCCTAC</td>
<td>ch r1</td>
<td>83830731</td>
<td>83830895</td>
<td>254</td>
</tr>
<tr>
<td>Un_Up_20 5 F1</td>
<td>GGGTACGTGTTATGGGTAGG</td>
<td>Un_Up_20 5 R</td>
<td>ATCTACCCACCTTCTACCCCTACCCCTAC</td>
<td>ch r1</td>
<td>97812534</td>
<td>97812573</td>
<td>219</td>
</tr>
<tr>
<td>Un_Up_20 5 F2</td>
<td>GGGTACGTGTTATGGGTAGG</td>
<td>Un_Up_20 5 R</td>
<td>ATCTACCCACCTTCTACCCCTACCCCTAC</td>
<td>ch r1</td>
<td>97812517</td>
<td>97812654</td>
<td>137</td>
</tr>
<tr>
<td>Un_Up_20 7 F</td>
<td>GGGTACGTGTTATGGGTAGG</td>
<td>Un_Up_20 7 R</td>
<td>ATCTACCCACCTTCTACCCCTACCCCTAC</td>
<td>ch r1</td>
<td>11179711</td>
<td>11179731</td>
<td>7</td>
</tr>
<tr>
<td>Un_UP_21 2 F1</td>
<td>AGAGAGTTAGGGAGTTAGGGAGG</td>
<td>Un_UP_212 R1</td>
<td>CCTAAAAAATCTTATAATATCCCTAC</td>
<td>ch r1</td>
<td>24844367</td>
<td>24844551</td>
<td>184</td>
</tr>
<tr>
<td>Un_UP_21 2 F2</td>
<td>GGTAGGGGATATTTTAAAGGTITTTTGAAGTTAGG</td>
<td>Un_UP_212 R2</td>
<td>ATCCCCRCAAACCTAACACC</td>
<td>ch r1</td>
<td>24844221</td>
<td>24844367</td>
<td>146</td>
</tr>
<tr>
<td>Un_UP_22 6 F</td>
<td>AGTTYAGGTTGGAGGATTTTTAGGTTG</td>
<td>Un_UP_226 R</td>
<td>AACCTTTCCAACCATTACAC</td>
<td>ch r1</td>
<td>76807217</td>
<td>76807414</td>
<td>197</td>
</tr>
<tr>
<td>Un_UP_22 9 F</td>
<td>AGTTTTAATAGGATYGTITTAGGG</td>
<td>Un_UP_229 R</td>
<td>TTCTRAACTACCTCACTAAAATTCACAA</td>
<td>ch r1</td>
<td>26025928</td>
<td>26026143</td>
<td>215</td>
</tr>
<tr>
<td>Un_UP_25 4 F1</td>
<td>GYGAGTTGGGGAAGAT</td>
<td>Un_UP_254 R1</td>
<td>CAACCCCAACCAACTTCC</td>
<td>ch r1</td>
<td>8847097</td>
<td>8847273</td>
<td>194</td>
</tr>
<tr>
<td>Un_UP_25 4 F2</td>
<td>GGAATTTTTGTTGGGTTG</td>
<td>Un_UP_254 R2</td>
<td>CACCAAATACCTATTTCTAACRCAC</td>
<td>ch r1</td>
<td>8847254</td>
<td>8847424</td>
<td>170</td>
</tr>
<tr>
<td>Un_UP_25 8 F</td>
<td>GYGTTAGGGGAGTGGAGT</td>
<td>Un_UP_258 R</td>
<td>CTCRCAAACCCCTTCTTAC</td>
<td>ch r1</td>
<td>40403223</td>
<td>40403436</td>
<td>213</td>
</tr>
<tr>
<td>Un_UP_26 2 F1</td>
<td>TTAYGGGAGTTGGGTTGAAA</td>
<td>Un_UP_262 R1</td>
<td>CTAACCCCRCTTACCACCA</td>
<td>ch r1</td>
<td>76066657</td>
<td>76066857</td>
<td>200</td>
</tr>
<tr>
<td>Un_UP_26 2 F2</td>
<td>GTTGGGTTAGTGGAGT</td>
<td>Un_UP_262 R2</td>
<td>CATTTCAACTTTTAAACCTCCCTAACAACC</td>
<td>ch r1</td>
<td>76066514</td>
<td>7606677</td>
<td>163</td>
</tr>
<tr>
<td>Un_UP_27 F</td>
<td>GTTYYGGGAGGYGGGAGTYYYGGGTTAGG</td>
<td>Un_UP_277 R</td>
<td>CCCCCCTAATCTACRCCCTTACACCAACC</td>
<td>ch r1</td>
<td>23210694</td>
<td>23210704</td>
<td>9</td>
</tr>
<tr>
<td>Un_UP_28 0 F</td>
<td>GGATTTGGGATTTGTTGGGG</td>
<td>Un_UP_280 R</td>
<td>CRTCCTTCCTTCTCTTTAAAACC</td>
<td>ch r2</td>
<td>4751448</td>
<td>4751638</td>
<td>190</td>
</tr>
<tr>
<td>Un_UP_28 1 F</td>
<td>GGATGGAAGTTGTTAGTGGT</td>
<td>Un_UP_281 R</td>
<td>CCCCTAAAAACTCAACTTCAATTCACTTAC</td>
<td>ch r2</td>
<td>21440888</td>
<td>21441099</td>
<td>211</td>
</tr>
<tr>
<td>Un_UP_28 3 F</td>
<td>GAGGGYGAGTTAGGTTAGGAT</td>
<td>Un_UP_283 R</td>
<td>ACRCACTTACCAACTTTACCTATACC</td>
<td>ch r2</td>
<td>24398359</td>
<td>24398524</td>
<td>165</td>
</tr>
<tr>
<td>Un_UP_28 6 F</td>
<td>AGTTGGGAGAGAATGATTYYGGG</td>
<td>Un_UP_286 R</td>
<td>ACRCTCCCTAATCCAACCTTCCACCA</td>
<td>ch r2</td>
<td>38749641</td>
<td>38749881</td>
<td>240</td>
</tr>
<tr>
<td>Un_UP_29 1 F</td>
<td>TTATTAGGTTTTTGTAGGTAAAT</td>
<td>Un_UP_291 R</td>
<td>CRRCRAATACRAAACCRRAAACAA</td>
<td>ch r2</td>
<td>50154089</td>
<td>50155014</td>
<td>205</td>
</tr>
<tr>
<td>Un_UP_29 3 F</td>
<td>TGGTTYCTTTTGTAGGAGTGAAGA</td>
<td>Un_UP_293 R</td>
<td>CTACCCCAATTCAACTTCCCTACACTTCCCATACC</td>
<td>ch r2</td>
<td>54638930</td>
<td>54639097</td>
<td>167</td>
</tr>
<tr>
<td>Un_UP_30 0 F</td>
<td>AATATTATTAGATGTTTCTATTAGGTCTTGGTTG</td>
<td>Un_UP_300 R</td>
<td>CCCCATATCTTACCAACTTAC</td>
<td>ch r2</td>
<td>27139087</td>
<td>27139241</td>
<td>154</td>
</tr>
<tr>
<td>Un_UP_30 6 F</td>
<td>GGYGTGGAAGGGTTYGATTATTTTG</td>
<td>Un_UP_306 R</td>
<td>CCTTATTTCTTCTTTAAACCTRACAAC</td>
<td>ch r2</td>
<td>29811183</td>
<td>29811342</td>
<td>159</td>
</tr>
<tr>
<td>Un_UP_30 7 F</td>
<td>GGGTVGGAGTGGTTTGGTAGT</td>
<td>Un_UP_307 R</td>
<td>CCACCCCCACCTTCCACACACAC</td>
<td>ch r2</td>
<td>38183127</td>
<td>38183316</td>
<td>189</td>
</tr>
<tr>
<td>Un_UP_31 6 F</td>
<td>GAGGAGTTGTAGGGGTTGAGTG</td>
<td>Un_UP_316 R</td>
<td>ACCACCTTTCCATCCACTCCTACTTC</td>
<td>ch r2</td>
<td>88846386</td>
<td>88846527</td>
<td>141</td>
</tr>
<tr>
<td>Un_UP_35 7 F</td>
<td>GATGGTGTTTGTGCCAATGGT</td>
<td>Un_UP_357 R</td>
<td>AAACRAATTCAAACCTACRAACCTAACTACT</td>
<td>ch r2</td>
<td>10030504</td>
<td>10030521</td>
<td>9</td>
</tr>
<tr>
<td>Un_UP_38 3 F</td>
<td>ATTTAGGGGTTTGGTGGTG</td>
<td>Un_UP_383 R</td>
<td>CTCAACACCCAAAACCTTACAACTTACCAACAC</td>
<td>ch r2</td>
<td>11563569</td>
<td>11563587</td>
<td>0</td>
</tr>
<tr>
<td>Un_UP_44 F</td>
<td>TGGTTAGAAGGAGTTGTTAGG</td>
<td>Un_UP_444 R</td>
<td>ATAATCTAAACTACCCAACTTTACAACTTACCAACTTACCAACTC</td>
<td>ch r2</td>
<td>22006990</td>
<td>22007002</td>
<td>7</td>
</tr>
<tr>
<td>Un_UP_45 F</td>
<td>GTTGGGAGTTGGGATTTGGGAT</td>
<td>Un_UP_454 R</td>
<td>CTCCACATCTTTCTGTTACGACCACTCAAC</td>
<td>ch r2</td>
<td>23306088</td>
<td>23306108</td>
<td>1</td>
</tr>
<tr>
<td>UnUp5-F1</td>
<td>GGTGGYGYGYGTGTTTTTTTGAGTATGTTT</td>
<td>UnUp5-R1</td>
<td>CAACRACRCCCTCCAACCTAACCRCR</td>
<td>ch</td>
<td>rl</td>
<td>14797760</td>
<td>14797911</td>
</tr>
<tr>
<td>-----------</td>
<td>---------------------------------</td>
<td>------------</td>
<td>-----------------------------</td>
<td>------</td>
<td>-----</td>
<td>---------</td>
<td>---------</td>
</tr>
<tr>
<td>UnUp5-F1a</td>
<td>GAATGGTGAGGCCTTTTTAATAAGAGGTGTTGG</td>
<td>UnUp5-R1</td>
<td>CAACRACRCCCTCCAACCTAACCRCR</td>
<td>ch</td>
<td>rl</td>
<td>14797724</td>
<td>14797911</td>
</tr>
<tr>
<td>UNUp5-F2</td>
<td>GGYGGAGGATATAAAGTGGAGGATGTTT</td>
<td>UnUp5-R2</td>
<td>CCTAACTAACCRCRATGACCRCCCACC</td>
<td>ch</td>
<td>rl</td>
<td>14797975</td>
<td>14798145</td>
</tr>
<tr>
<td>Un_Up_52</td>
<td>GAATTGYGTTTCTTTGTTTAAGATGTTT</td>
<td>Un_Up_52</td>
<td>TACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r3</td>
<td>12759629</td>
<td>12759650</td>
</tr>
<tr>
<td>Un_Up_59</td>
<td>GGATTTGGGATTTAGGGATAGGATG</td>
<td>Un_Up_59</td>
<td>TTACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r3</td>
<td>12759629</td>
<td>12759650</td>
</tr>
<tr>
<td>Un_Up_59</td>
<td>GATTTGGGGATAGGATAGGATG</td>
<td>Un_Up_59</td>
<td>TTACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r3</td>
<td>12759629</td>
<td>12759650</td>
</tr>
<tr>
<td>Un_Up_62</td>
<td>TYGATTGGTTTCTTTGTTTAAGATGTTT</td>
<td>Un_Up_62</td>
<td>TACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r3</td>
<td>12759629</td>
<td>12759650</td>
</tr>
<tr>
<td>Un_Up_66</td>
<td>GAGGATTTAGGGATTTAGGGATAGGATG</td>
<td>Un_Up_66</td>
<td>TACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r3</td>
<td>12759629</td>
<td>12759650</td>
</tr>
<tr>
<td>Un_Up_76</td>
<td>GGAAGGTGTTAAGAATGITTGGTTT</td>
<td>Un_Up_76</td>
<td>TTACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r5</td>
<td>76285711</td>
<td>76285893</td>
</tr>
<tr>
<td>Un_Up_96</td>
<td>GGTGGAGGATATT&gt;TATGTTGGGT</td>
<td>Un_Up_96</td>
<td>TACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r6</td>
<td>73388590</td>
<td>73388705</td>
</tr>
<tr>
<td>Un_Up_98</td>
<td>GTATTTAAGGGATTTTGTAGTGG</td>
<td>Un_Up_98</td>
<td>TACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r6</td>
<td>73388590</td>
<td>73388705</td>
</tr>
<tr>
<td>Un_Up182</td>
<td>GAGAAGGATATGTTAGGAGGATA</td>
<td>Un_Up182</td>
<td>TACCCAAAACCTCRAACCCAAACA</td>
<td>ch</td>
<td>r1</td>
<td>61079374</td>
<td>61079566</td>
</tr>
</tbody>
</table>

*Hg18 Coordinates*
Appendix Table 2. Window Primer Sequences.

<table>
<thead>
<tr>
<th>Oligo Name</th>
<th>Sequence (5' to 3')</th>
</tr>
</thead>
<tbody>
<tr>
<td>Un_up_106_3FM</td>
<td>GGCGtAGAGTttCGtAGAGG</td>
</tr>
<tr>
<td>Un_up_106_3FU</td>
<td>GGtGtAGAGTtttGtAGAGG</td>
</tr>
<tr>
<td>Un_up_106_3RM</td>
<td>CCACGACGGAATCTTAACCTAC</td>
</tr>
<tr>
<td>Un_up_106_3RU</td>
<td>CCACAACACAAACTTAACCTAC</td>
</tr>
<tr>
<td>Un_up_146_2FM</td>
<td>TtAttTTGGGtTTGGTGttTG</td>
</tr>
<tr>
<td>Un_up_146_2FU</td>
<td>TtAttTTGGGtTTGGTGttTG</td>
</tr>
<tr>
<td>Un_up_146_2RM</td>
<td>TATTCCCTACGAACTCCTTC</td>
</tr>
<tr>
<td>Un_up_146_2RU</td>
<td>TATTCCCTACAAACTCCTTC</td>
</tr>
<tr>
<td>Un_up_146_4FM</td>
<td>TGGGtTTGGTGttTG</td>
</tr>
<tr>
<td>Un_up_146_4FU</td>
<td>TGGGtTTGGTGttTG</td>
</tr>
<tr>
<td>Un_up_146_4RM</td>
<td>TAACGCTCCCCTATCCCTAC</td>
</tr>
<tr>
<td>Un_up_146_4RU</td>
<td>TAACACTCCCCTATCCCTAC</td>
</tr>
<tr>
<td>Un_up_207_5FM</td>
<td>CAACCCCCGCTCCCCAAACCCAAAC</td>
</tr>
<tr>
<td>Un_up_207_5FU</td>
<td>CAACCCCCGCTCCCCAAACCCAAAC</td>
</tr>
<tr>
<td>Un_up_207_5RM</td>
<td>GAGACGAAGAGAGGCGGTGAGAA</td>
</tr>
<tr>
<td>Un_up_207_5RU</td>
<td>GAGAtGAAGAGAGgtGGTGAGAA</td>
</tr>
<tr>
<td>Un-Up_307_1aFM</td>
<td>tCGCGCGttCGtCCGCGTTttTtt</td>
</tr>
<tr>
<td>Un-Up_307_1aFU</td>
<td>tCGCGCGttCGtCCGCGTTttTtt</td>
</tr>
<tr>
<td>Un-Up_307_1aRM</td>
<td>CCCCACTTCCACCCCGTCC</td>
</tr>
<tr>
<td>Un-Up_307_1aRU</td>
<td>CCCCACTTCCACCCCATCC</td>
</tr>
</tbody>
</table>

U denotes primer specific for unmethylated DNA. M denotes primer specific for methylated DNA.
REFERENCES


81. Singhi AD, Davison JM, Choudry HA, Pingpank JF, Ahrendt SA, Holtzman MP, et al. GNAS is frequently mutated in both low-grade and high-grade disseminated appendiceal mucinous neoplasms but does not affect survival. Hum Pathol. 2014;45: 1737–1743. doi:10.1016/j.humpath.2014.04.018


doi:10.1016/S0002-9440(10)65346-1

doi:10.1016/j.gde.2012.08.001


doi:10.1053/j.gastro.2014.10.041

doi:10.1038/ng.2796

doi:10.1038/ng.2591


146. Souza RF. The role of acid and bile reflux in oesophagitis and Barrett’s metaplasia. Biochem Soc Trans. 2010;38: 348. doi:10.1042/BST0380348


