G, WAS GOING ON? PUTATIVE REGULATORY FUNCTION OF
GWAS-IDENTIFIED MARKERS OF SUSCEPTIBILITY TO ACUTE APPENDICITIS

by

Ekaterina Orlova

B.A., French Literature and B.S., Molecular Biology,

University of California at San Diego, 2008

Submitted to the Graduate Faculty of the

Department of Human Genetics

Graduate School of Public Health in partial fulfillment

of the requirements for the degrees of

Master of Public Health and Master of Science

University of Pittsburgh

2015

UNIVERSITY OF PITTSBURGH

Graduate School of Public Health

This thesis was presented

by

Ekaterina Orlova

It was defended on

August 7, 2015

and approved by

Committee Member
David N. Finegold, M.D.
Professor, Human Genetics, Graduate School of Public Health
Professor of Medicine and Pediatrics, School of Medicine
University of Pittsburgh

Academic Advisor & Committee Co-chair
John R. Shaffer, Ph.D.
Assistant Professor, Human Genetics, Graduate School of Public Health,
University of Pittsburgh

Thesis Advisor & Committee Chair
Michael J. Morowitz, M.D., FACS
Assistant Professor of Surgery, University of Pittsburgh School of Medicine,
Attending Pediatric Surgeon, Division of Pediatric General and Thoracic Surgery,
Children's Hospital of Pittsburgh of UPMC

Michael Morowitz, MD, FACS

John Shaffer, PhD

G, WAS GOING ON? PUTATIVE REGULATORY FUNCTION OF
GWAS-IDENTIFIED MARKERS OF SUSCEPTIBILITY TO ACUTE APPENDICITIS

Ekaterina Orlova, M.S., M.P.H

University of Pittsburgh, 2015

ABSTRACT

Appendicitis affects 7-9% of Americans and is the most common diagnosis requiring hospitalization of both children and adults. Several etiologies of appendicitis have been hypothesized, but definitive mechanisms remain elusive – a critical review of the literature does not support a primary role of fecaliths or lymphoid hyperplasia, as is commonly believed. It is known that appendicitis has heritable components, and so we collaborated with 23andMe Inc., a personal genomics company, to identify genetic determinants of susceptibility to acute appendicitis. 23andMe performed a genome-wide association study (GWAS) of 18,773 appendectomy cases and 114,907 controls, and identified one locus with genome-wide significance. In addition, the GWAS identified eight highly significant SNPs that did not reach genome-wide significance. Most of the SNPs identified using this analysis fell outside of protein-coding genes, thus bioinformatic analysis using RegulomeDB was done to interrogate the SNPs' putative regulatory capacity of nearby or distant genes, or proteins.

This analysis identified 921 targets of putative regulatory elements in the same LD blocks as the four of nine lead SNPs identified in the GWAS and chosen for follow-up study. Of these, 299 targets were unique when targets from all four genomic regions were combined. These targets were organized according to the distance of their putatively

regulatory SNP from the given lead SNP, and based on overlap of elements' targets within one region with targets of elements within the rest of the genomic regions. Ultimately, the following list of 17 proteins was generated for priority in further studies: CEBPB, CTCF, EP300, EVI-1, FOS, FOXJ3, FOXP1, GATA1, HNF4A, JUN, MYC, NFKB, PPARG, RAD21, SPI1, STAT1, and STAT3. This list includes several proteins that directly interact with, or influence the expression of very specific inflammatory markers known to be strongly associated with appendicitis, including IL-8, IL-1B, and IL-6. This outcome supports the utility of RegulomeDB in the interpretation of GWAS-generated non-coding variants.

This compiled resource and the ongoing parallel studies born of the appendectomy GWAS may help to elucidate the pathogenesis of acute appendicitis, thereby providing opportunities to improve the diagnosis, treatment, and prevention of this extremely common disease. The public health significance of appendicitis and its genetics are addressed, and a theoretical public health program that integrates the multiple factors involved in appendicitis etiology is proposed.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

PREFACE

I would like to extend enormous thanks to my committee members, and all of the individuals who contributed their expertise during the many meetings that brought this project to fruition: Dr. Michael Morowitz, Dr. John Shaffer, Dr. David Finegold, Dr. Michael Barmada, Dr. Robert Ferrell, Dr. David Whitcomb, and Dr. Candace Kammerer. I am extremely grateful for their encouragement, indispensable scientific guidance, and especially the ongoing commitment to the project in spite of the many life hardships members our group experienced in the last two years. I consider it a privilege to have worked with you all.

Thank you to 23andMe and its research participants for providing the crucial data that enabled our follow-up studies, and especially to Dr. Carrie Northover for working with our group to establish this important collaboration. I would also like to thank Dr. Robin Grubs for her very key and timely thesis-related guidance. A big thank you to the individuals with IT and programming knowledge who have been enormously helpful in managing the 4 million (!) lines of data from RegulomeDB - Andrew Lau, Maxim Strukov, and my dad. Thank you also to Dr. Alan Boyle for his RegulomeDB insight. A special thanks to Noel Harrie for her tireless assistance over the last two years, and to Joanne Pegher for her help with ETD preparation. Thank you also to Kamal Patel for the initial information regarding 23andMe's Research Portal.

Finally, heartfelt thanks to my family for their support, to Anusha, to all the lovely cafes of Pittsburgh, and especially to Andrew for most everything else.

1.0   INTRODUCTION

Appendicitis is a common, complex disease of unknown etiology. In this study, the genetic component of the disease was investigated through a multidisciplinary collaboration between a public university and a private company. The basis for the study was citizen science - the sizeable genetic data set analyzed was gathered from consenting research participants who had purchased a direct-to-consumer genetic test and had answered medical history questions.

In this work, a genome-wide association study (GWAS) of appendectomy identified one single nucleotide polymorphism (SNP) with genome-wide significance, and eight SNPs that were highly significant, but did not reach this threshold. As expected, most of these nine lead SNPs fell outside of protein-coding genes, and thus their role in appendicitis etiology was less easily understood than if they had clearly altered the structure or function of known genes. It is possible that lead SNPs that fall within non-coding regions are instead involved in the regulation of distant or nearby genes that contribute to disease pathogenesis or are protective of it by this less direct mechanism. If these identified non-coding SNPs represent true associations, it is imperative to use available resources to interpret their significance for the disease. One such novel resource is RegulomeDB[1] – a database of putative regulatory variants, the genetic elements by which the variants are thought to exert their effects, and the elements' target genes or proteins. RegulomeDB has previously been used to successfully interrogate GWAS associations falling in protein coding and non-coding regions[2]. The broad aim of this study was to use this

database to help interpret the role of the non-coding variants identified in the appendectomy GWAS.

The first specific aim was to prioritize which SNPs associated with appendectomy to examine in this follow-up study. A second aim was to extract from RegulomeDB the putative regulatory variants from the vicinity of the prioritized lead SNPs, and to organize them based on the level of evidence supporting their capacity to be regulatory. A third aim was to further prioritize these SNPs, the regulatory elements by which they are thought to exert their effects, and the gene or protein targets of these elements for follow-up study. A final aim was to develop hypotheses as to the putative role of some of the targets identified in appendicitis etiology based on existing literature.

2.0   BACKGROUND

2.1   APPENDICITIS AND ITS ETIOLOGY

Acute appendicitis is the inflammation of the vermiform appendix, a tubular organ that protrudes at the base of the cecum. It frequently presents with central abdominal pain, followed by vomiting and migration of the pain to the right-inferior part of the abdomen[3]. The initial pain is colicky in nature for the first 24 hours, and transforms to a more constant and severe pain after migration. Loss of appetite, constipation, and nausea co-occur often[4].

Acute appendicitis is considered a common disease; it affects 9.38 per 10,000 people in the United States every year[5]. It is the fifth most common indication for non-neonatal pediatric hospitalization, and is the second most common inpatient pediatric procedure[6]. Incidence peaks among children aged 10-14; it is more common in males than females, and more common among whites and Hispanics, relative to other races. The lifetime risk of developing appendicitis is 9%[7].

The appendix is commonly thought to be evolutionarily vestigial in humans, left over from a time when our ancestors were herbivorous, and the organ was longer and served as a reservoir of cellulose-digesting bacteria[8]. It has also been proposed to be an immune organ due to the presence of significant gut-associated lymphoid tissue (GALT) within it[9]. More recent research suggests that the appendix may play a key role in the maintenance of commensal intestinal bacteria that, in turn, play a large role in the human immune system[10,11]. Findings that

3

support this assertion include the appendix's unique position in the intestinal tract, its shape, the abundance of mucus production within it, and that lymph tissue supports mutualistic biofilms in the gut[12].

Interestingly, there is increasing evidence that appendicitis results from a complex interaction between host genotype and the microbial environment of the intestines (the gut microbiome). Expression and protein studies of inflamed and non-inflamed appendices have found altered mRNA expression and differences in protein levels within the enterocytes, the cells lining the intestines. DMBT1, a secreted glycoprotein thought to play a role in enterocyte differentiation and bacterial defense, was found to have five-fold increased expression and corresponding increases in protein level in inflamed appendices[13]. In addition, several studies have shown that appendicitis is associated with the local growth of the gram negative pathogen, *Fusobacterium nucleatum*[14-16]. In a series of 52 inflamed appendices from several countries, 62% had invasion of this pathogen[15]. This organism is not typically present at significant levels in the GI tract; it is normally found in the oral cavity. It is a central player in the etiology of periodontal disease, and its translocation is frequently found in extra-oral infections, including the amniotic fluid of pre-term infants[17]. The bacterial risk factors for appendicitis may be set in place from a young age. In one study, children with appendicitis were found to have been breastfed for 74% of the duration of the time of children without appendicitis[18]; breastfeeding is known to modify infant immune response to microbial agents.

The leading theory as to the cause of appendicitis states that the organ's inflammation results from an obstructive process of the lumen – the tubular cavity of the appendix. The obstruction is thought to be due to fecaliths (hard masses of feces) or lymphoid hyperplasia[19]. However, several studies have not supported fecaliths nor lymphoid hyperplasia as a primary

cause of obstruction in a majority of cases of appendicitis[20-23]. Instead, they suggest that local inflammation precedes the luminal obstruction[24]. In addition, dietary intake of fiber is considered to be a significant factor in appendicitis etiology; lower intake of all fiber fractions is frequently found among appendicitis patients relative to controls[25]. Finally, the "hygiene hypothesis" has been proposed as part of the explanation for appendicitis etiology, namely that improved sanitation practices in industrialized countries have resulted in individuals having less exposure to microbes. This is thought to result in corresponding "over-reactions" to later infections that then trigger appendicitis[26]. Barker *et al*, proponents of the hygiene hypothesis, argued against a dietary cause of appendicitis by citing the example of the blacks in South Africa who – in spite of eating a Westernized, low-fiber diet – had low rates of appendicitis[27].

An alternative proposed etiology of appendicitis is that it is precipitated by Type 1 Hypersensitivity, a type of allergic response, and that infection is a later consequence[28]. A finding that supports this theory is that the levels of eosinophils – pro-inflammatory leukocytes equipped to participate in gastrointestinal tract inflammation – have been found to be significantly elevated in the serum of individuals with appendicitis, relative to controls[29]. Eosinophils have also been found to be significantly elevated within the muscularis of the appendix (the smooth muscle surrounding the appendix) in acute appendicitis along with features of mast cell degranulation, relative to control appendices. This was suggested to be an early finding in the pathogenesis of the disease – not a consequence of subacute or chronic inflammation[28]. Eosinophils have also been found to be elevated in the irritable bowel disease ulcerative colitis, but their definitive role in disease remains elusive[30]. However, a subsequent study of inflammatory gene expression of inflamed appendices found a very focused

inflammatory response and concluded that appendicitis was unlikely to be due to a Type 1 or Type 2 immune response[31].

Finally, Ballester *et al* found that individuals who had had a tonsillectomy had a 2.57-fold increased odds for subsequent appendectomy[32]. The authors proposed two possible explanations. First, tonsillectomy could produce a deficiency in lymphoid tissue which induces the GALT tissue within the appendix to overcompensate in response to incoming pathogens and become overly inflamed. Alternatively, higher rates of tonsillectomy and appendectomy could both be a result of genetic predisposition to a hyperactive immune response and hypertrophy.

Although the first official diagnosis of appendicitis and subsequent appendectomy were performed in 1880[33], a specific etiology of acute appendicitis has not been firmly established to this day.

2.2 APPENDICITIS DIAGNOSIS AND ITS TREATMENT

The diagnosis of acute appendicitis depends primarily on clinical findings. There is no diagnostic test for appendicitis, but urinalysis and serum screens can help rule out some two dozen differential diagnoses. The Alvarado score is a clinical scoring system occasionally used to aid in diagnosis, and depends on the patient's medical history, physical examination, and blood lab tests[4]. The imaging techniques ultrasonography and computed tomography (CT) scanning are also used to aid in diagnosis, however a longitudinal study of these techniques did not find that their introduction lowered false positive diagnoses resulting in unnecessary appendectomies ("negative appendectomies")[34]. Other studies have shown more favorable results for CT scanning. However, with CT scanning there is concern regarding unnecessary exposure of long

duration to the risks of ionizing radiation[4]. Unfortunately, these technologies are not available at all hospitals, and when they are, making arrangements for their use can further delay diagnosis. Thus, there remains a need for a rapid, non-invasive diagnostic test for the condition.

The treatment of choice for appendicitis is the surgical removal of the appendix – appendectomy – within the first 24 hours of the onset of symptoms[35]. Delays past this time window are associated with an increased risk of perforation of the appendix, and spillage of the contents of the appendix into the abdominal cavity. This results in a worse prognosis and a longer hospital stay. If surgery is delayed more than 36 hours after the onset of symptoms, the rate of perforation can be as high as 36%[36]. Broad spectrum antibiotics are typically administered to help prevent postoperative wound infections and intra-abdominal abscesses.

2.3   GENETICS OF APPENDICITIS

Epidemiological and genetic studies of acute appendicitis suggest that it is a complex, multifactorial disease with environmental, bacterial, and several genetic components. Currently, there is no evidence for a single gene cause of appendicitis[37]. However, the existence of a rare single genetic cause in a small proportion of cases can't be ruled out – there have been reports of families with up to 39 individuals affected, some of whom shared anatomic defects of the organ[38]. Nevertheless, the vast majority of cases are presumed to be due to multifactorial causes, and mathematical models predict that the genetic component is polygenic[37].

Acute appendicitis demonstrates clear heritability in family and twin studies[37,39]. Genetics account for 30-56% percent of the risk of appendicitis; and environmental effects account for the remainder of the risk[37,40]. Other studies have found the heritability to be 27% and the effect of

7

shared familial environment to be 16%[41]. A positive family history of acute appendicitis increases the relative risk of appendicitis up to 10-fold[9]. A number of association studies have been done between polymorphisms in genes involved in innate immunity and the inflammatory response, and the occurrence or severity of appendicitis. C-allele carriage at −174 in the IL-6 gene (rs1800795) is associated with severe appendicitis, and with lower plasma and peritoneal fluid IL-6 protein levels[42]. However, this study examined a small number of SNPs, and the sample size was limited.

2.4   GENE EXPRESSION IN APPENDICITIS

It is known that appendicitis is correlated with differential expression of a core set of genes involved in the inflammatory response, with similar genes being activated in cases of mild and severe appendicitis. This set of genes is highly enriched in mediators of the innate inflammatory response and is specific to acute appendicitis (as distinct from other inflammatory diseases of the bowel)[31]. The specific gene expression profile of acute appendicitis lends support to a bacterially-mediated etiology of the disease. In particular, one cytokine gene differentially expressed in acute appendicitis (IL-1β) is strongly induced by bacterial products such as lipopolysaccharide (LPS, a part of the outer membrane of gram-negative bacteria) [31]. In addition, the study also found a significant upregulation of the neutrophil chemoattractant interleukin-8 (IL-8) in proportion with the extent of inflammation in the appendix[31]. IL-8 is induced through several pathways and in response to LPS, TNF, IL-1, and through cell-mediated immunity. It has also been found to have elevated expression in inflammatory bowel diseases[43-45].

IL-8 upregulation in peritoneal fluid of individuals with appendicitis has been reported previously[46,47]. Interestingly, in an assay screening various pathogenic bacteria isolated from peritoneal exudate fluids of patients with appendicitis, *Fusobacterium necrophorum* was capable of induction of IL-8 from cultured human mesothelial cells to levels found *in-vivo* in peritoneal fluids of patients with appendicitis. This bacterium when heat-killed, and its supernatant also induced elevated IL-8[47].

Serum levels of various interleukins have been studied as possible diagnostic markers. In one study, IL-6, but not IL-8 expression has been found to be elevated in the serum of patients with appendicitis[48]. In another, both pre-operative IL-6 and IL-8 were found to be higher in patients with perforated as compared with non-perforated appendicitis[49].

2.5   THE GENETICS OF COMMON COMPLEX DISEASE

Common complex diseases are ones that do not display a clearly recognized inheritance pattern and typically have several factors that contribute to their etiology like environmental exposures such as diet, infectious disease, toxins, and internal factors like aging.

There exist several theories for the role genetics plays in complex diseases that are common in the general population. The common disease, common variant (CD/CV) hypothesis states that the genetic contribution to diseases common in the general population - the heritability - is moderated by a combination of several common genetic variants. It follows that due to the high frequency of these variants, their individual contributions to the overall disease risk – their effect sizes - are small.

Common risk alleles are thought to be common because their negative effect on relative fitness is small, and thus selection pressure against them is correspondingly small. It is likely that these risk alleles became common many thousands of years ago, in an environment that differed significantly from the one that exists today. Indeed, the prevalence of appendicitis is greater in industrialized nations than in developing ones, and explanations of this phenomenon have been in line with the evolutionary mismatch hypothesis (EMH)[26,27,50]. The EMH states that disease common in the general population is due to a mismatch between the environment that humans had evolved in, and the pressures of the modern lifestyle; certain evolved traits may be maladaptive and lead to disease in the context of modern civilization.

There are many additional factors that could have produced a random increase in risk alleles. For example, it is also possible that a variant that increases the risk for a certain disease is, in fact, protective against other ailments. Alternative explanations for the genetics of common complex disease include the common disease, rare variant (CD/RV) hypothesis, which states that common disease is caused by multiple rare variants with larger effect sizes. The determination of which of these hypotheses applies more to appendicitis would greatly influence the clinical applications of the genetic variants identified – for example, therapeutic and prognostic assays would be greatly simplified if only a handful of variants were known to influence disease risk[51].

The genome wide association study (GWAS) is a genetic technique for identifying common genetic variants that are associated with a given disease in a group of individuals. A GWAS strives to take a broad, un-biased "discovery" approach to determining the genetic determinants of disease by querying thousands of SNPs across the genome at once. This can be contrasted with a more limited approach of focusing research on a select few genes that are predicted to be most important to disease etiology given their function. The strength of the

GWAS study is in numbers: by comparing large sample sizes of individuals with and without disease (typically in the thousands), it is able to identify the genetic differences of small effect size associated with the particular disease. Given the CD/CV hypothesis, this approach is particularly powerful for elucidating the more "subtle" genetics of common complex disease like acute appendicitis.

The output data of a GWAS study are SNPs (single nucleotide polymorphisms) or other genetic variants that are associated with increased or decreased risk for the condition. However, identifying SNPs in a GWAS is only a preliminary step in identifying risk loci for a common disease such as appendicitis. This is so because SNPs identified through GWASs don't necessarily "cause" disease, but are markers for other genetic factors that may truly modify risk -- often, SNPs have been inherited within sections of DNA for generations (termed linkage disequilibrium), and thus their association with a disease is in reality a surrogate for the true risk factor in their vicinity. An ideal situation for the interpretation of GWAS results is when the associated SNP is causal – for example, it is located inside a gene which makes the said gene function sub-optimally and results in an increased likelihood of disease. However, this scenario is rare. Given that the vast majority of our DNA (~97%) is not made up of protein-coding genes, it is not surprising that most SNPs identified in GWASs fall in intergenic regions. These regions likely have regulatory activity of nearby or distant sites[2], thus to truly understand their significance, it is necessary to examine the identified polymorphisms' regulatory potential.

The aim of the present study was to interpret the significance of the SNPs identified in the GWAS of appendectomy conducted by the company 23andMe. Specifically, the goal was to use RegulomeDB to identify the variants with putative regulatory potential within the regions in the same LD block as the lead SNPs of the GWAS. Once identified, a list of the elements

putatively regulated by these variants would be compiled, and the regulatory targets would be organized and prioritized to enable more focused follow-up study.

The variants identified as being potentially regulatory through this study were screened for their regulatory effects on three types of elements: eQTLs, motifs, and proteins (transcription factors). Expression quantitative trait loci (eQTLs) are typically polymorphisms that significantly influence the expression of near or distant genes. Protein binding sites are genetic sequences that have an affinity for certain proteins to bind, typically with the aim of inducing or repressing transcription. The strength of this affinity can be modified by regulatory polymorphisms within it or at more distant locations. Motifs are sequences of several nucleotides in length that can be binding sites for transcription factors, which, in turn, regulate expression of nearby genes. Of note regarding nomenclature: the "targets" of motifs listed in this study do not refer to these nearby genes – instead, motif "targets" identified using RegulomeDB refer to the transcription factors whose binding is affected by the putative regulatory variant within the motifs.

3.0   SIGNIFICANCE

3.1   ACUTE APPENDICITIS AND PUBLIC HEALTH

Acute appendicitis affects 9% of females and 7% of males in the United States[19]. More than 300,000 appendectomies are performed annually[52]. Mortality due to non-perforated appendicitis is 0.8 in 1,000, but 5.1 in 1,000 for perforated appendicitis[4]. Complications can include wound infection (up to 20%, if appendix is perforated), and intra-abdominal or pelvic abscesses. Recovery time can range from 10 to 28 days in adults, depending on complications and individual risk factors, however hospital stays are between 2 and 5 days long, on average[53].

Disparities exist in time to diagnosis, and those presenting later are at a higher risk of having perforated appendicitis. This is especially true for young children and elderly individuals. Perforation rates at presentation to the hospital in young children can be as high as 97%[4], in comparison with an average rate of approximately 30%[53]. Individuals with schizophrenia are also at increased risk of having perforated appendicitis, largely due to delays in time to diagnosis[54]. In addition, the following other groups have higher rates of perforation and longer time to presentation to the healthcare system: African Americans, individuals covered by Medicaid, and those who are uninsured[53].

An appendectomy for non-perforated appendicitis has an average cost of $7,800. An average appendectomy for perforated appendicitis costs 50% more - $12,800[53]. Hospital charges

for appendectomies can range from \$1,529 to \$182,955, with a median price of \$33,000 for uncomplicated appendicitis[55].

Certain populations are at greater risk of complications due to appendicitis. Appendicitis during pregnancy can be fatal in 4% of mothers in cases of perforated appendicitis; 1 in 1,000 pregnant women are affected by acute appendicitis[56]. The fetus is also at 20-35% risk of death in cases of perforated appendicitis[57]. The diagnosis of appendicitis is particularly complicated during pregnancy because the fetus' as well as the mother's wellbeing must be taken into account when weighing the risks and benefits of the various diagnostic procedures. In these situations, laparoscopic diagnosis of appendicitis can result in a negative appendectomy in a staggering 40% of cases, and radiation exposure to the fetus and the mother during imaging must be taken into account[56].

Interestingly, appendectomy can relieve or prevent other gastrointestinal conditions. Appendectomy significantly improves the symptoms of the large majority (90%) of patients with ulcerative proctitis, a subtype of ulcerative colitis. In 40% of these patients, the surgery leads to complete remission of symptoms[58]. Furthermore, a meta-analysis has shown that appendectomy reduces the risk of developing ulcerative colitis by 67%[59].

Although there have been great improvements in the diagnosis of appendicitis – in previous decades the negative appendectomy rate exceeded 20% of all appendectomies[60] – approximately every one in twelve appendectomies is unnecessary today[61]. Women are more than twice as likely as men to have a negative appendectomy[61]. Negatives appendectomies are not without medical consequences: they are fatal in approximately 1% of cases[61].

4.0   METHODS

4.1   MULTIDISCIPLINARY COLLABORATION

The University of Pittsburgh team submitted a research proposal to receive access to de-identified aggregate analysis from 23andMe database of genotype and phenotypic data from research participants. The data received from 23andMe included the results of the GWAS conducted by 23andMe on the appendectomy phenotype. Lead SNPs identified through the GWAS study were subsequently annotated by the University of Pittsburgh group using RegulomeDB and organized for follow-up study. The University of Pittsburgh team's collaboration with 23andMe, the team's access to the GWAS data and supporting documents from 23andMe's database, and the subsequent annotation of the GWAS data for this project were approved by the Institutional Review Board (IRB) of the University of Pittsburgh (**Appendix B**).

4.2   GWAS STUDY DESIGN

A genome-wide association study was conducted by 23andMe using the appendectomy phenotype on data from 23andMe research participants, who provided informed consent to participate in research under the 23andMe research protocol approved by the external AAHRPP-

accredited IRB, Ethical & Independent Review Services (E&I Review). This cohort of research subjects has been described previously[62-65].

4.3   STATISTICAL ADJUSTMENT

Logistic regression was performed assuming an additive model for allelic effects using the model: *appendectomy ~ age + sex + pc.0 + pc.1 + pc.2 + pc.3 + pc.4 + genotype.* The age variable refers to age at the time of genotyping, not age at appendectomy. Covariates for age, gender, and the top five principal components of ancestry were included to account for residual population structure. The genomic control procedure was used to compensate for variance inflation due to residual population stratification that had not been effectively controlled for through use of principal components in the regression models. The results were adjusted for a calculated genomic control inflation factor of 1.034.

4.4   GENOTYPING AND QUALITY CONTROL MEASURES

The GWAS included research participants' genetic data generated through the use of the 23andMe® Personal Genome Service (PGS), a saliva-based direct-to-consumer genotyping service. 23andMe has used four genotyping platforms since it released the PGS in 2007, and data from these four platform were included in the GWAS - The V1 and V2 platforms were variants of the Illumina HumanHap550 BeadChip with additional custom SNPs curated by the 23andMe

research team. The V3 platform is a variant of the Illumina OmniExpress+ BeadChip, also with custom SNPs. The current V4 platform is a custom array.

The GWAS also included imputed SNPs computed against the March 2012 "v3" release of 1000 Genomes haplotypes (phase 1 variants list)[66]. Data from each genotyping platform was phased and imputed separately. Beagle[67] (version 3.3.1) was used to phase batches of 8000-9000 individuals across chromosomal segments of 10,000 or fewer genotyped SNPs, with overlaps of 200 SNPs. The following SNPs were excluded: those with call rate < 95%, with Hardy-Weinberg equilibrium $P<10^{-20}$, or with large allele frequency discrepancies compared to European 1000 Genomes reference data. The phased segments were imputed against 1000 Genomes haplotypes of all ethnicities (excluding monomorphic and singleton sites); a high-performance version of Minimac[68] was used, with 5 rounds and 200 states for parameter estimation.

Males and females were phased together in segments for the non-pseudoautosomal region of the X chromosome. Males were treated as already phased, while the pseudoautosomal regions were phased separately. Next, males and females were imputed together using minimac, as was done with the autosomes, treating males as homozygous pseudo-diploids for the non-pseudoautosomal region.

For tests using imputed data, the imputed dosages were used, rather than the best-guess genotypes. The association test P value reported was computed using a likelihood ratio test. Results for the X chromosome were computed similarly – men were coded as if they were homozygous diploid for the observed allele.

17

4.5   POPULATION DESCRIPTION

The GWAS was restricted to research participants who had more than 97% European ancestry; ancestry was determined through comparison with the three HapMap2 populations[69]. Close relatives were excluded using a segmental identity-by-descent (IBD) estimation algorithm[70]. Close relatives were defined as those who share more than 700 cM IBD - either one or both genomic segments IBD - which corresponds to the amount of expected sharing between first cousins in an outbred population (approximately 20% of genetic information).

4.6   APPENDECTOMY PHENOTYPE DETERMINATION

The appendectomy phenotype for the GWAS was ascertained based on research participants' voluntary answers to online health history questionnaires deployed in the 23andMe website. The appendectomy cases were identified from two questions in two separate questionnaires: "Have you ever had your appendix removed?"; answer choices consisted of "yes," "no," and "I'm not sure." The second question was "Have you ever had any of the following *other* surgeries?"; answer choices to the "appendectomy" selection included "yes," "no," and "I don't know." Cases answered in the affirmative to either question, while controls answered in the negative. Individuals who responded with discordant results to the two questions were excluded from the study.

4.7   ASSOCIATION ANALYSIS

For quality control of genotyped GWAS results, SNPs were flagged with a Hardy-Weinberg $P<10^{-20}$ in Europeans, a minor allele frequency of <0.1%, or a call rate of <90%. SNPs that were only genotyped on the 23andMe V1 platform were also flagged, due to limited sample size. Genotyped SNPs were also tested for date effects, and flagged SNPs with $P<10^{-50}$ by ANOVA of SNP genotypes against a factor dividing genotyping date into 20 roughly equal-sized buckets.

For imputed GWAS results, SNPs with avg.rsq<0.5 or min.rsq<0.3 in any imputation batch were flagged, as well as SNPs that had strong evidence of an imputation batch effect. The batch effect test is an F test from an ANOVA of the SNP dosages against a factor that represents imputation batch; results with $P<10^{-50}$ were flagged. Prior to performing the GWAS, the largest subset of the data passing these criteria was identified for each SNP, based on its original genotyping platform (either v2+v3+v4, v3+v4, v3, or v4 only) and association test results were computed for whichever was the largest passing set. Consequently, there were no imputed results for SNPs that failed these filters. Across the merged results of genotyped and imputed SNPs, logistic regression results that did not converge due to complete separation were flagged.

4.8   FUNCTIONAL ANNOTATION OF LEAD SNPS USING REGULOMEDB

Select lead SNPs from the GWAS and regions within the same LD blocks were further annotated for variants with putative functional significance using RegulomeDB[1] (version 1.1, publicly available at regulome.stanford.edu). RegulomeDB is a database which guides interpretation of human regulatory variants and includes data from the ENCODE Project[71] and more than 962

other sources. It has recently been expanded further to stay up-to-date with current ENCODE releases[72,73] as well as Chromatin States from the Roadmap Epigenome Consortium (unpublished) and several other updates. The database employs a scoring system (on a scale 1-6) which helps filter for variants most likely to be regulatory or with demonstrated regulatory function. The scoring system is further detailed in **Table** 1**.**

Table 1. RegulomeDB Scoring System.

Score	Supporting Data Types
1a	eQTL + TF binding + matched TF motif + matched DNase Footprint + DNase peak
1b	eQTL + TF binding + any motif + DNase Footprint + DNase peak
1c	eQTL + TF binding + matched TF motif + DNase peak
1d	eQTL + TF binding + any motif + DNase peak
1e	eQTL + TF binding + matched TF motif
1f	eQTL + TF binding / DNase peak
2a	TF binding + matched TF motif + matched DNase Footprint + DNase peak
2b	TF binding + any motif + DNase Footprint + DNase peak
2c	TF binding + matched TF motif + DNase peak
3a	TF binding + any motif + DNase peak
3b	TF binding + matched TF motif
4	TF binding + DNase peak
5	TF binding or DNase peak
6	Other

Listed are the score and the corresponding data types available to support the assertion of regulatory potential. Variants with scores 1a-1f are likely to affect binding and linked to expression of a gene target. Variants with scores 2a-2c are likely to affect binding. Variants with scores 3a-3b are less likely to affect binding. Variants with scores 4-6 represent minimal evidence of binding."Other" data types represent more rare forms of evidence.

Lead SNPs were prioritized for further annotation based on the *p*-value of the lead SNP, and the relative density of the most highly associated SNPs in the vicinity of each lead SNP. In regions that did not have nearby LD peaks, the boundaries for demarcating the regions of interest were also established based on density of associated SNPs. In total, four genomic regions were

selected for the annotation study: chromosome 3 (49,360,000–50,100,000 bp), chromosome 15 (73,240,000–73,640,000 bp), chromosome 4 (111,610,058–111,737,533 bp), and chromosome 4 (112,755,000–112,895,000 bp).

Within these regions, all variants were examined for putative regulatory function, defined as a score of 3 or less (regulome.stanford.edu, accessed [14 Dec 2014] and updated [29 Mar 2015]). If a variant with a score of 3 or less fell within a motif, its role within the motif was judged to be significant if its location had at least 33% conservation; and the element was retained for analysis. This degree of conservation was approximated visually.

5.0   RESULTS

5.1   GWAS IDENTIFIES ONE LOCUS WITH GENOME-WIDE SIGNIFICANCE

A genome-wide association study of appendectomy was conducted on genomes of 133,680 individuals; the data were filtered to remove close relatives and included only individuals of >97% European ancestry. The appendectomy phenotype was ascertained based on answers to online questionnaires regarding appendectomy; cases answered "yes," they had had an appendectomy (n=18,773), and controls answered "no" (n=114,907).

The demographics of the study population are shown in **Table 2.** The proportion of cases to controls among the European cohort studied is 16.3% to 83.7%, and is in line with approximate expected rates of appendicitis (14%) in urban whites in the United States in 1979[74], and the prevalence of appendicitis in European countries, including over the last thirty years in Greece (16.4%)[74], and during 1960-1965 in the UK (15-18%)[75].

The GWAS study results were adjusted for age, sex, and the top five principal components of ancestry[1]. The Q-Q plot of the p-values is available in **Figure 7.**

---

(1) [1] The genome-wide association study data showed a female gender bias – women were more likely than men to report having had appendectomy (B= 0.32097; $P = 1.8\times10^{-86}$). Since there is a male bias in true appendicitis, it is likely that our data set reflects an excess of women who have had incidental appendectomy. This is in line with national statistics: prior to 1990, women aged 35-44 had a 12.1-fold increased risk of incidental appendectomy (43.8 per 10,000 population per year) relative to men. Regardless, the GWAS was corrected for gender.76. Addiss DG, Shaffer N, Fowler BS, Tauxe RV. The

One locus met genome-wide significance (p-value $< 5 \times 10^{-8}$): rs2129979 (p-value 8.8×10^{-14}). Additionally, eight single nucleotide polymorphisms (SNPs) were highly significant, but did not reach the genome-wide significance threshold: rs192656182 (p-value 9.5×10^{-8}), rs137882920 (p-value 9.9×10^{-8}), rs2247036 (p-value 1.0×10^{-7}), rs17044095 (p-value 3.2×10^{-7}). rs117367662 (p-value 5.3×10^{-7}), rs1650337 (p-value 6.9×10^{-7}), rs75972139 (p-value 7.8×10^{-7}), and rs6445791 (p-value 9.6×10^{-7}). Lead SNP rsIDs are reported based on the snp137CodingDbSNP schema from the UCSC Genome Browser. These nine index SNPs are described in **Table 4.** The results are also displayed graphically in the Manhattan Plot in **Figure 1**. Quality statistics for the index SNPs are shown in **Table 3.** Regional association plots for the index SNPs and surrounding regions included in the RegulomeDB annotation analysis are shown in **Figure 2, Figure 3, Figure 4,** and **Figure 5.** Regional association plots for index SNPs not included in the annotation analysis are shown in **Figure 8, Figure 9, Figure 10, Figure 11,** and **Figure 12.**

Table 2. Demographics of unrelated, European individuals included in Appendectomy GWAS.

Phenotype	Group	Total	Male	Female	Age 0-30	Age 30-45	Age 45-60	Age 60+
Appendectomy	**Case**	18,773	8,175	10,598	763	2,702	4,823	10,485
	Control	114,907	59,824	55,083	14,984	33,077	31,761	35,085

The ages shown represent the age at which patients were genotyped**.**

epidemiology of appendicitis and appendectomy in the United States. *American journal of epidemiology.* Nov 1990;132(5):910-925.
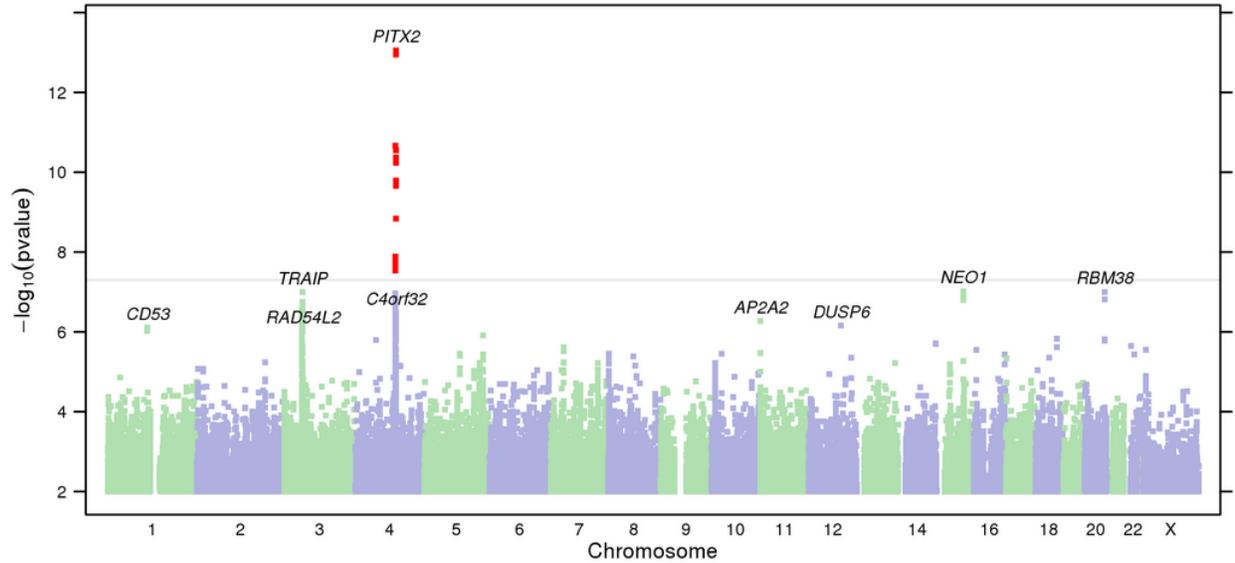
Figure 1. Manhattan Plot of GWAS of Appendectomy.

The nine index SNPs, shown annotated with names of nearest genes, are depicted as a distribution of association test statistics versus genomic position. The Y-axis depicts the −log10-transformed P-values from the association test; the X-axis depicts the chromosomes 1-22, X and Y. The grey line represents a p-value of 5 X 10^{-8}, and the result surpassing this threshold is shown in red.

Table 3. Index SNPs for Strongest Genome-Wide Associations.

Chromosome	SNP rsID	Position	Allele	P value	OR (95% CI)	SNP control/case	Gene context
4	rs2129979	111,720,997	G/T	8.8×10^{-14}	0.908 (0.886,0.932)	0.7116/ 0.6943	*PITX2* --- []
15	rs192656182	73,599,970	C/T	9.5×10^{-8}	1.453 (1.272,1.661)	0.008/ 0.0104	*NEO1* – [] --*HCN4*
20	rs137882920	56,005,258	C/T	9.9×10^{-8}	0.745 (0.666,0.833)	0.0178/ 0.0146	*RBM38*--[]--*CTCFL*
3	rs2247036	49,882,349	C/T	1.0×10^{-7}	1.065 (1.040,1.090)	0.4731/ 0.4884	[TRAIP]
4	rs17044095	112,777,414	G/T	3.2×10^{-7}	1.074 (1.045,1.104)	0.7680/ 0.7787	[] ---C4orf32
11	rs117367662	967,822	C/T	5.3×10^{-7}	0.863 (0.815,0.915)	0.0603/ 0.0543	[AP2A2]
12	rs1650337	89,770,068	G/T	6.9×10^{-7}	Inf	0.0000/ 0.0013	DUSP6--[]---GALNT4
1	rs75972139	111,373,721	A/G	7.8×10^{-7}	0.800 (0.733,0.872)	0.9843/ 0.9810	KCNA3---[]--CD53
3	rs6445791	51,601,982	A/G	9.6×10^{-7}	1.084 (1.050,1.120)	0.1660/ 0.1758	[RAD54L2]

Reported are the most-associated SNPs within each associated region for appendectomy cases and controls. Gene context graphically depicts the distance between the index SNP ("[]") and the nearest gene; " "= <1kb, "−" = <10kb, "--" = <100kb, "−−−" = <1000kb. Inf stands for infinity.
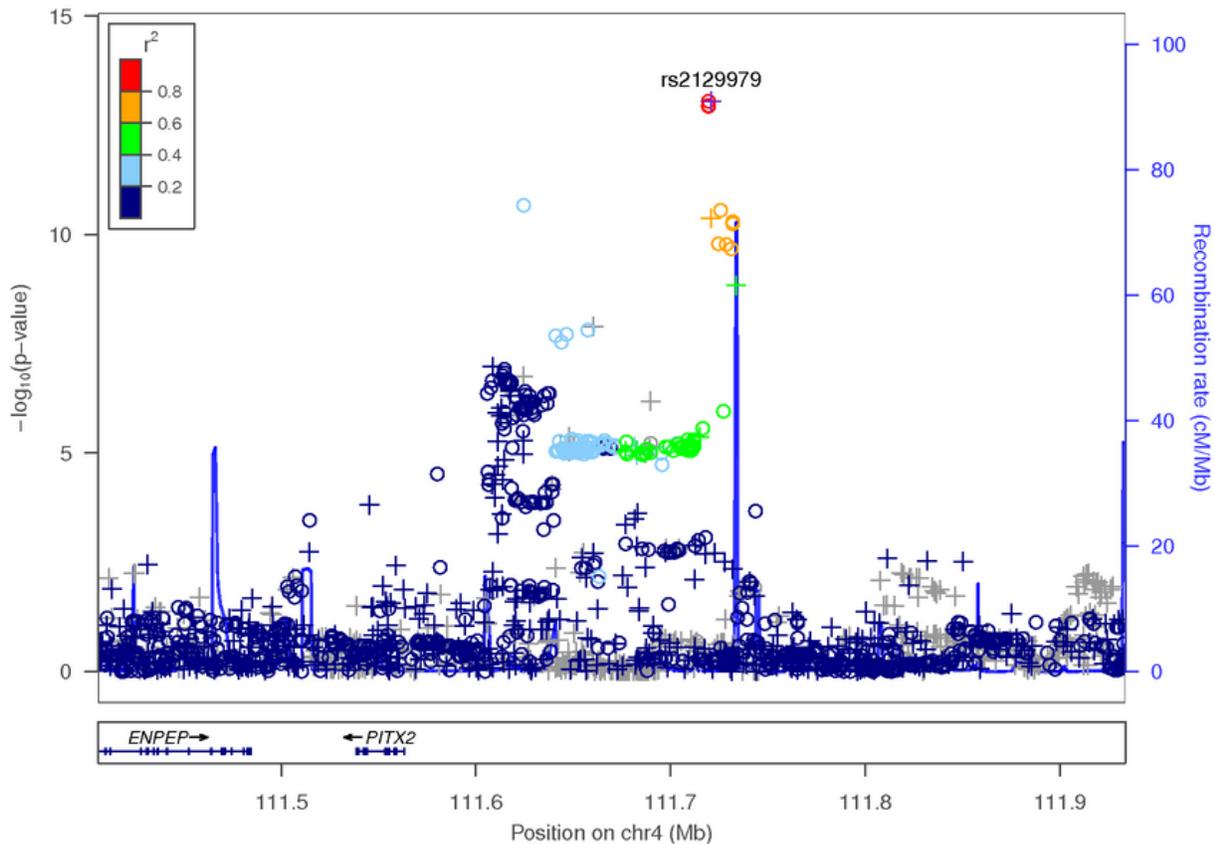
Figure 2. Regional Association Plot for rs2129979.

Association test results are shown as a distribution of position on chromosome 4 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. Genes in the vicinity of the lead SNP are shown along the X-axis. The region 111,610,058–111,737,533 bp was chosen for further analysis in the RegulomeDB annotation study.
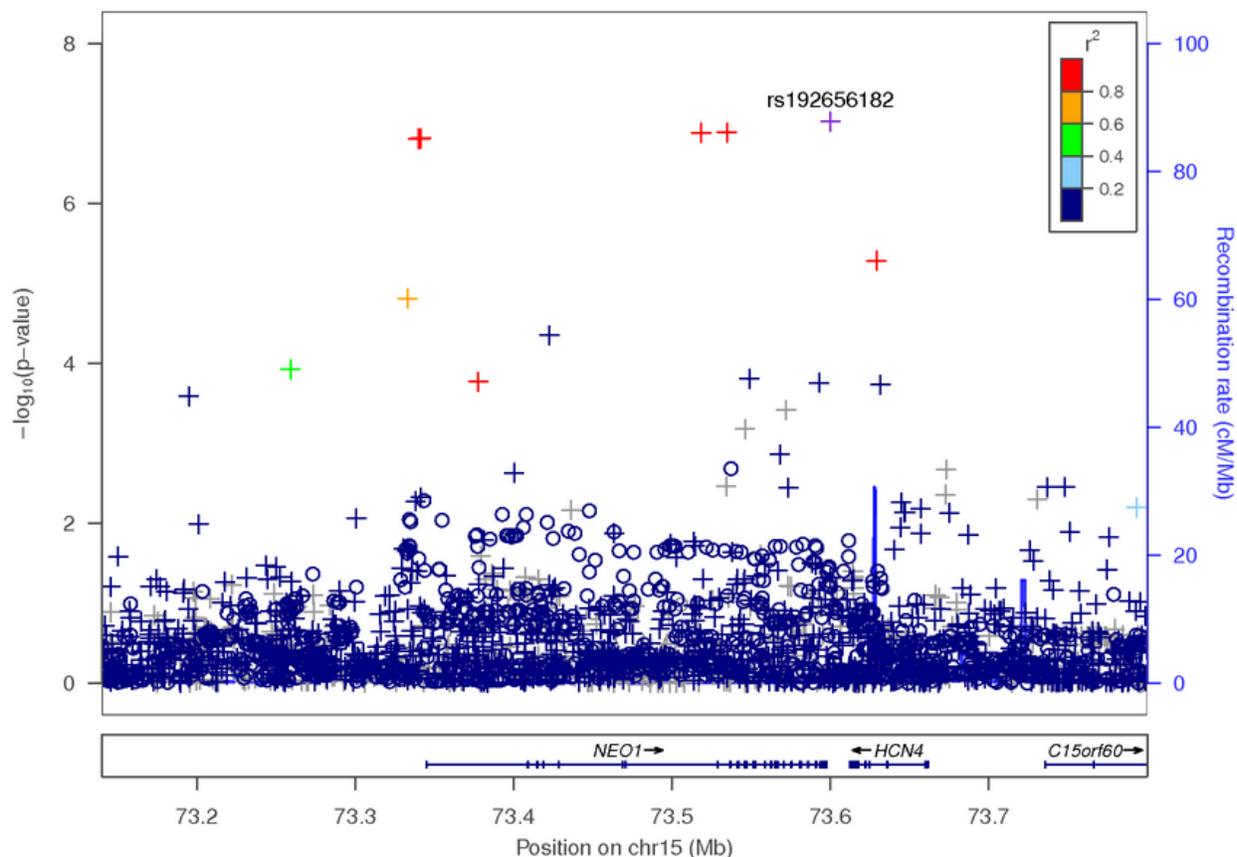
Figure 3. Regional Association Plot for rs192656182.

Association test results are shown as a distribution of position on chromosome 15 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. Genes in the vicinity of the lead SNP are shown along the X-axis. The region 73,240,000–73,640,000 bp was chosen for further analysis in the RegulomeDB annotation study.
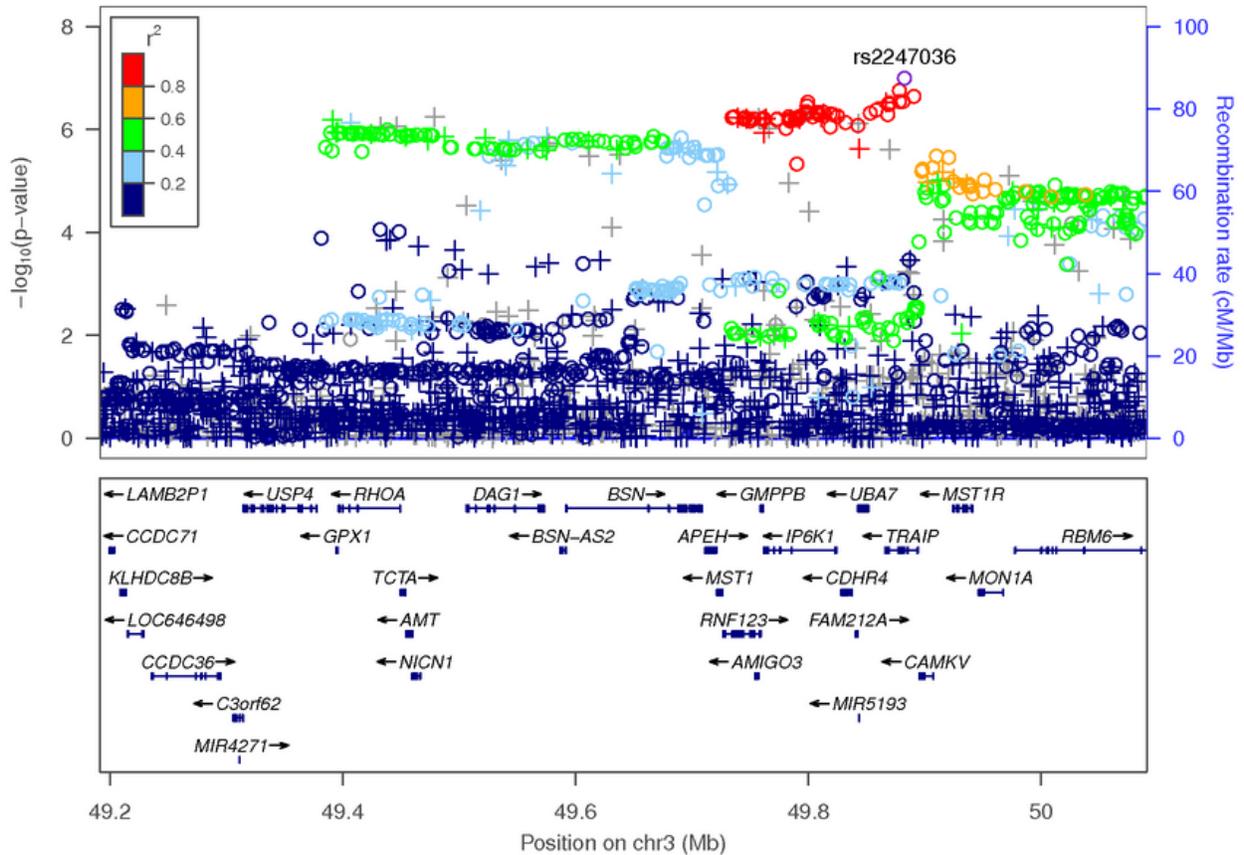
Figure 4. Regional Association Plot for rs2247036.

Association test results are shown as a distribution of position on chromosome 3 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. Genes in the vicinity of the lead SNP are shown along the X-axis. The region 49,360,000–50,100,000 bp was chosen for further analysis in the RegulomeDB annotation study.
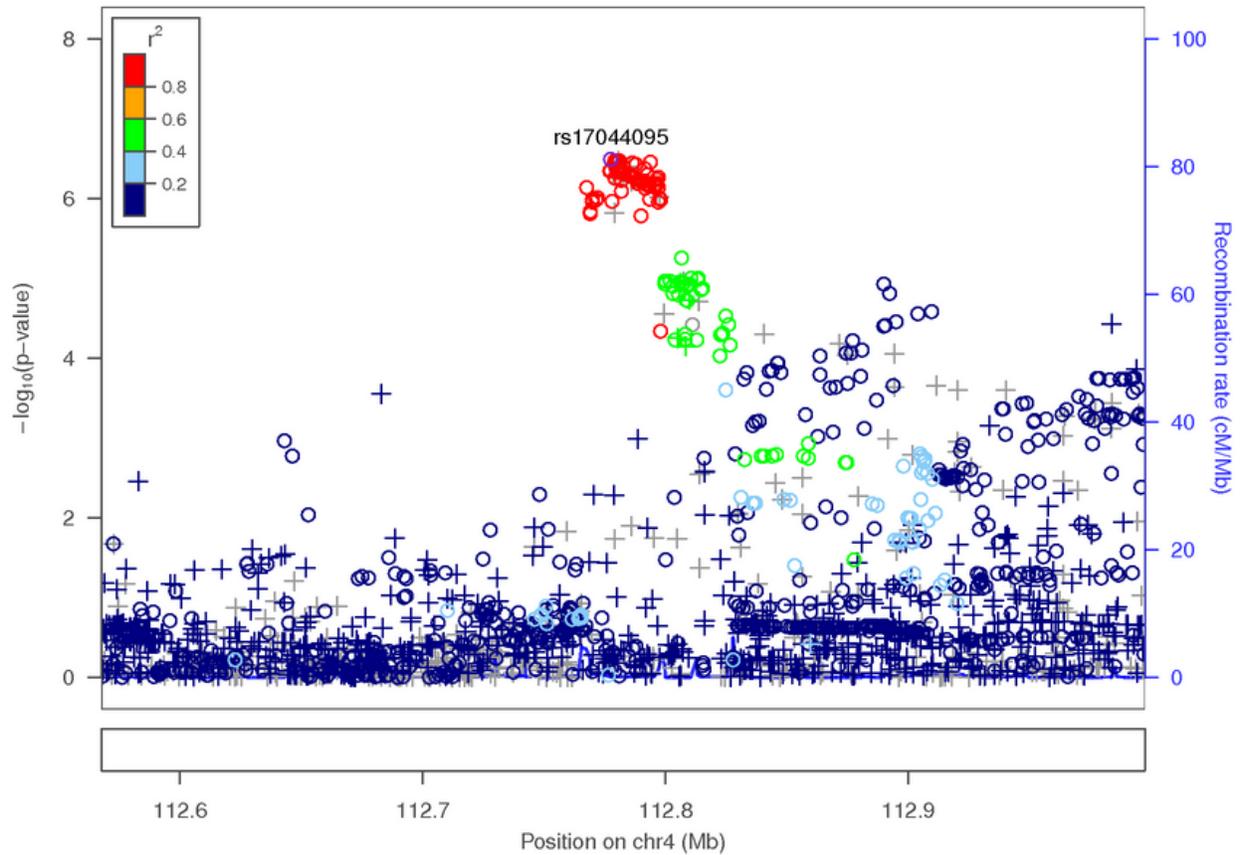
Figure 5. Regional Association Plot for rs17044095.

Association test results are shown as a distribution of position on chromosome 4 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. The region 112,755,000–112,895,000 bp was chosen for further analysis in the RegulomeDB annotation study.

5.2 IDENTIFICATION OF 299 TARGETS PUTATIVELY REGULATED BY ELEMENTS ASCERTAINED USING REGULOMEDB

The genomic regions examined using RegulomeDB identified a total of 4,579 putative regulatory variants. The region on chromosome 15 (73,240,000–73,640,000 bp) returned 1,507 SNPs with regulatory potential, of which 40 SNPs had a score of 3 or less, indicating higher potential for being regulatory. This region yielded 109 unique targets (eQTL gene targets, protein binding targets, or proteins with affinity for specific motifs) potentially regulated by these SNPs.

The region on chromosome 4 (111,610,058–111,737,533 bp) returned 576 SNPs, of which 6 SNPs had a RegulomeDB score of 3 or less, and yielded 13 targets. The region on chromosome 3 (49,360,000–50,100,000 bp) returned 1,903 SNPs, of which 132 SNPs had a score of 3 or less, and yielded 254 unique targets[2]. The second region on chromosome 4 (112,755,000–112,895,000 bp) returned 593 SNPs, of which 21 SNPs had a score of 3 or less, and yielded 24 unique targets.

In total, 921 targets of potentially regulatory elements were identified using RegulomeDB (**Table 4**). The motifs identified along with their targets are organized by chromosomal region and available in **Table 5** (chromosome 3), **Table 6** (chromosome 15), **Table 7** (chromosome 4 at 111.7 Mb), and **Table 8** (chromosome 4 at 112.7 Mb). None of the nine lead SNPs of the GWAS were found to have any putative regulatory activity.

---

[2] This region on chromosome 3 also yielded two dsQTLs (DNase I sensitivity quantitative trait loci), both with a score of 1f. dsQTL stands for "DNase I sensitivity QTL" – a location at which DNase sequencing read depth correlates significantly with a nearby SNP or insertion/deletion. dsQTLs are typically enriched in predicted transcription factor binding sites, and are associated with transcription factor binding changes that are allele-specific77. Degner JF, Pai AA, Pique-Regi R, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature.* Feb 16 2012;482(7385):390-394.

Many targets of the targets of the identified elements overlapped. Of the 921 targets, 299 targets were unique across all four chromosomal regions examined. Seven targets were found to be putatively regulated by variants within three of four chromosomal regions: *EP300, EVI-1, FOXJ3, FOXP1, GATA1, HNF4A*, and *POLR2A*. Four were targeted by variants within all four chromosomal regions: *CEBPB, CTCF, FOS*, and *RAD21*.

Elements were further stratified based on the distance of their associated putative regulatory variant's distance from the index SNP. Details for elements found near chromosome 3 are listed in **Table 9**, those for chromosome 15 in **Table 10**, those for chromosome 4 (at 111.7 Mb) in **Table 11**, and those for chromosome 4 (at 112.7 Mb) in **Table 12**.

6.0   DISCUSSION

To the best of my knowledge, this was the first GWAS study of appendectomy performed. It identified one SNP far surpassing genome-wide significance (rs2129979). Four of the lead SNPs from the study and the surrounding regions within the same LD blocks were then further annotated for the existence of putative functional variants using RegulomeDB (rs2129979, rs192656182, rs2247036, rs17044095).

The RegulomeDB work identified 299 unique targets of regulatory elements modified by variants in the same LD block as the lead SNPs of the GWAS. In addition, four of these targets had putative regulatory input from variants in all four regions examined (CEBPB, CTCF, FOS, RAD21), and seven targets had putative regulatory input from variants in three of the four regions (EP300, EVI-1, FOXJ3, FOXP1, GATA1, HNF4A, POLR2A). POLR2A is the B1 subunit of RNA Polymerase II; given that there are several protein-coding genes in the vicinity of the lead SNPs in all four regions, it is unsurprising that polymerase may be a target of the putative regulatory elements in the region. There was also overlap between the targets putatively regulated by variants across several chromosomal regions and targets putatively regulated by variants which were found nearest to the lead SNPs - within 10 kb: CTCF, RAD21, and within 50 kb: CEBPB, CTCF, FOS, RAD21, EP300, GATA1, HNF4A, and POLR2A.

6.1 IDENTIFICATION OF TARGETS THAT INTERACT WITH KNOWN KEY PLAYERS IN THE INFLAMMATORY RESPONSE IN APPENDICITIS: IL-8, IL-1B AND IL-6

Several targets of putative regulatory elements were found to interact with known inflammatory markers of appendicitis, and will be described below. CEBPB (CCAAT/Enhancer Binding Protein, beta) is a transcription factor with a key role in the regulation of genes involved in the immune and inflammatory responses[78]. According to the RegulomeDB analysis, it is putatively regulated by rs576813 which is 11,588 bp from the lead SNP on chromosome 15, with a score of 2b ("likely to affect binding"). It is also putatively regulated by rs9814765 which is 63,794 bp from the lead SNP on chromosome 3, with a score of 2c ("likely to affect binding"); there are additional SNPs with putative regulatory activity of CEBPB in this region of interest farther than this. CEBPB is also putatively regulated by rs7434417 which is 12,559 bp from the lead SNP on chromosome 4 (at 111.7 Mb), with a score of 3a ("less likely to affect binding"). Finally, CEBPB is putatively regulated by rs7569015, located 98,642 bp from the lead SNP on chromosome 4 (at 112.7 Mb), also with a score of 3a.

FOS is a regulator of cell proliferation, differentiation, and transformation, and has roles in stress response and apoptosis[79-81]. It is putatively regulated by rs2252833 which is 121,860 bp from the lead SNP on chromosome 3 with a score of 1d ("likely to affect binding and linked to expression of a gene target"), along with several other SNPs at more distant locations within this region of interest. It is also putatively regulated by rs576813 which is 11,588 bp from the lead SNP on chromosome 15, with a score of 2b ("likely to affect binding"), along with several SNPs at more distant locations within this region of interest. It is putatively regulated by rs7434417 and rs7439625, which are 12,559 bp and 57,519 bp away from the lead SNP on chromosome 4

(at 111.7 Mb) with scores of 3a for both ("less likely to affect binding"). Finally, it is putatively regulated by rs7289990, rs7674382, and rs757507, which are 65,676 bp, 67,788 bp and 105,786 bp away from the lead SNP on chromosome 4 (at 112.7 Mb) with scores of 3a, 2b and 3a. According to StringDB (a protein-protein interaction database[82] that has been used previously to prioritize genes identified in GWAS studies[83]), CEBPB interacts with other proteins from the RegulomeDB data set: PPARG, SPI1, and MYC.

IL-8 is the predominant chemokine/cytokine that is upregulated in mild appendicitis[31]. IL-8 is also the only chemokine/cytokine that is upregulated in severe appendicitis, although IL-1 and IL-11 are also differentially expressed in severe vs mild appendicitis. Because its expression is upregulated in both severe and mild appendicitis, it is predicted that IL-8 plays a central role in appendicitis pathogenesis[31]. Interestingly, according to StringDB, both CEBPB and FOS interact directly with interleukin 8 (IL-8). JUN and NFKB1, two other proteins within the RegulomeDB data set, also interact directly with IL-8 (FOS and JUN can work together as part of the inducible transcription complex AP-1).

It has been reported that in acute appendicitis, a very targeted innate immune response is mounted that includes Interleukin-1, beta (IL-1B, a pro-inflammatory cytokine), but not TNF, although both are strongly induced in response to products of bacteria such as LPS[31]. According to StringDB, of the 10 displayed proteins that IL-1B interacts directly with, three are encoded by genes represented in the RegulomeDB analysis: FOS, JUN, and NFKB1.

Serum levels of interleukin-6 (IL-6) have been found to be elevated in individuals with appendicitis, however IL-6 was not one of the interleukins upregulated in a comprehensive study of inflammatory gene expression in inflamed appendices relative to controls. According to

StringDB, of the 10 displayed proteins IL-6 interacts directly with, 6 are represented within the RegulomeDB data set: CEBPB, FOS, JUN, NFKB1, STAT1, and STAT3.

NFKB1 (nuclear factor kappa, beta subunit 1) and NF-IL6 (nuclear factor interleukin-6, also known as CEBPB) are known to synergistically activate transcription of the inflammatory cytokines implicated in appendicitis, IL-6 and IL-8[84]. In addition, although C/EBP alone only weakly binds to the IL-8 promoter, together with NFKB, it displays synergism and cooperativity in binding to this promoter. The regulation of IL-8 expression depends on the ratio of cellular C/EBP and NFKB[85]. Virtually all pathways that result in upregulation of IL-8 also indirectly elevate AP-1 (made up of FOS/JUN), and NFKB[86].

As was demonstrated above, most of the targets identified after filtering the RegulomeDB analysis results for close variant distance and overlap across multiple regions showed evidence of interacting with the very specific inflammatory pathways known to be activated in acute appendicitis. Thus, this preliminary confirmation of molecular players involved in appendicitis demonstrates that the data collected through RegulomeDB show promise as a tool for uncovering additional insights into the etiology of appendicitis on a molecular level.

Many variants studied in this work have low RegulomeDB scores, and correspondingly higher likelihood of regulating the genetic elements compiled, and so the entirety of the data can be used as a resource for further genetic analysis. However, based on the preceding protein-protein interaction findings, the following priority list of proteins is proposed for more immediate follow-up studies: CEBPB, CTCF, EP300, **EVI-1**, FOS, **FOXJ3**, **FOXP1**, GATA1, HNF4A, JUN, MYC, NFKB, **PPARG**, RAD21, SPI1, **STAT1**, and STAT3. Bolded proteins are ones that were identified based on their association with a motif, while unbolded proteins are

those whose binding is affected by putative regulatory elements. A flowchart of the process of narrowing down relevant targets from the initial 4,579 regulatory variants is available **Figure 6.**
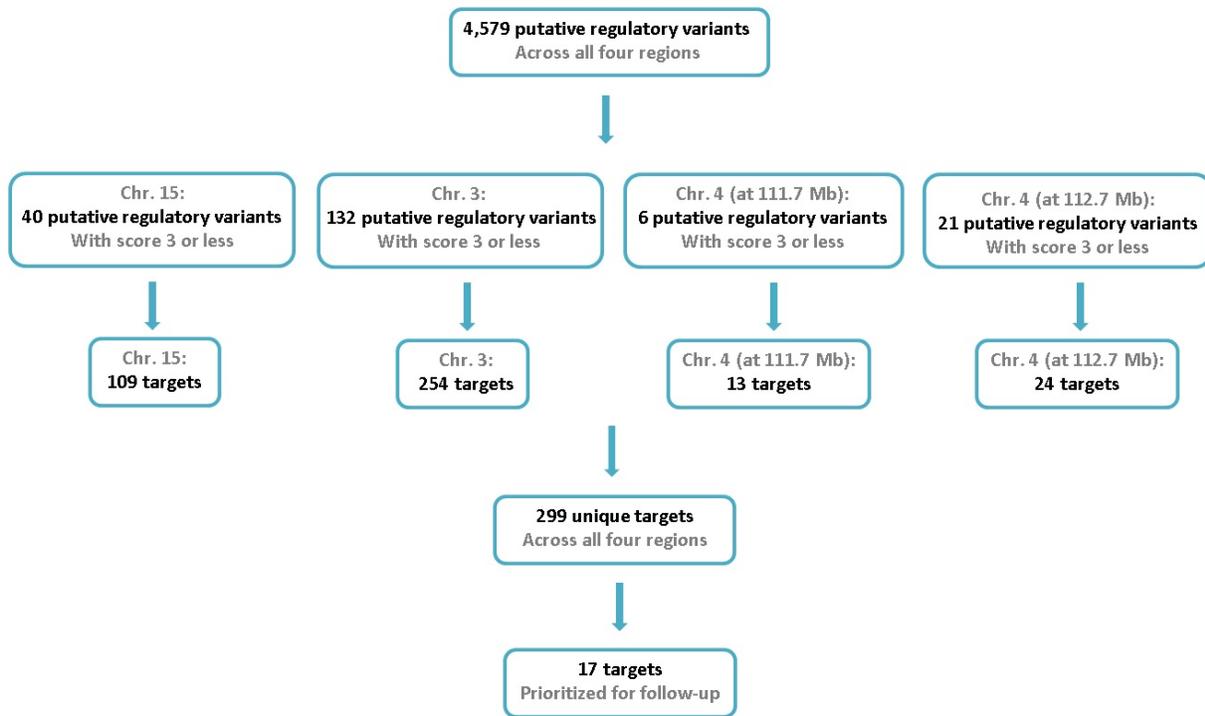
Figure 6. Flowchart of RegulomeDB Variant and Target Filtering Process.

6.2   LIMITATIONS OF THE CURRENT STUDY

There are several limitations to the GWAS study. The appendectomy phenotype is based on self-report by individuals, and has not been confirmed with medical records. Thus it is not possible to guarantee that all cases of appendectomy were, in fact, true appendectomies. Also, the appendectomy outcome is not a perfect substitute for appendicitis. Before the more sensitive technologies that aid in the diagnosis of appendicitis today were developed, there were many more negative appendectomies for every true case of appendicitis. Thus it is possible that the

"false" cases of appendicitis are skewing the GWAS signals identified. However, the rate of correct diagnosis has greatly improved over time: currently, one incidental appendectomy is performed for every 9 true cases of appendicitis (Michael Morowitz, M.D., FACS, personal communication).

To assess the potential impact of this limitation with respect to the GWAS findings, one can examine the age variable in the GWAS – the age at which an individual got genotyped. Because appendicitis primarily affects individuals aged 10-19, and this pattern hasn't changed in recent decades[5], one might suppose that in general, individuals who were older at the time of genotyping had their appendix removed earlier in time than individuals who were younger at the time of genotyping. One might then expect that if there was a strong confounding effect of age on the data set that there would be a stronger association between the lead SNPs of the GWAS and appendectomy in the younger age group at the time of genotyping, given that these individuals were more likely to have "true appendicitis," and a less strong association within older age groups. Similarly, individuals who were younger at the age of genotyping may have been able to more accurately recall the diagnosis of appendicitis (versus cholecystectomy, for example) because they would have had it, on average, more recently than those genotyped at older ages. However, the associations identified were not significantly stronger in any particular age group, which lends further support to the validity of the SNPs identified. Another limitation is that this study was done in a >97% European population, thus the findings in this study may not be representative of appendicitis in a different ethnic group.

There are also several limitations to the RegulomeDB follow-up study. The data regarding regulatory effects of polymorphisms found within RegulomeDB has only been verified in certain cell types, and may not translate to other cell types. The three types of regulatory

mechanisms queried in this study are not exhaustive of all types of regulatory elements, thus the regions of interest examined may have other regulatory potential. Finally, examining the role of protein-coding genes within the regions of interest was outside the scope of this study. Indeed, some of the variants identified in the GWAS, and SNPs in the same LD block with them, do fall within protein coding genes, thus it is possible that the SNPs identified have a more direct role in the pathogenesis besides regulating a secondary target. Identifying the role of these protein-coding genes is one of the future research directions discussed below.

6.3   CONCLUSIONS

To interpret the appendectomy GWAS findings, RegulomeDB was used to identify a list of target genes and proteins putatively regulated by variants in the vicinity of the four prioritized lead SNPs of the GWAS. A starting list of 299 unique targets identified based on likelihood of being regulated by SNPs in the regions queried was further filtered based on two factors. First, the targets were organized based on their putative regulatory variants' distance from the prioritized lead SNPs of the GWAS. This increased the likelihood of the variants identified as being potential sources of the GWAS signal. Second, the targets were filtered based on whether they were putatively regulated by variants across three or four of the four genomic regions identified queried. The resulting identification of genes encoding several interaction partners of inflammatory factors known to have strong associations with appendicitis lends support to the validity and potential of this independent discovery method.

6.4   FUTURE RESEARCH DIRECTIONS

The immediate goal of future studies will be to replicate the SNPs correlated with increased or decreased risk of developing appendicitis. A follow-up collaborative replication study involving the lab of Dr. Michael Morowitz of Children's Hospital of Pittsburgh of UPMC and the University of Cincinnati Children's Hospital that genotypes appendicitis subjects and controls at the SNPs identified in the GWAS study is ongoing.

An additional means of advancing this compiled resource of priority SNPs from RegulomeDB is to integrate it with other publicly available bioinformatics tools for annotation of non-coding variants, such as dbPSHP, CADD, and GWAVA. This method has been shown to be very effective for further refinement of a list of priority SNPs related to irritable bowel disease (IBD)[87]. In addition, as mentioned previously, the motif targets identified in this study refer to the transcription factors which are affected by putative regulatory elements within the motifs. A follow-up study might examine the genes whose expression is modified by the motifs identified.

An ongoing study in the lab of Dr. Michael Morowitz is examining the gene expression of inflamed and non-inflamed appendices, as well as examining the serum of appendicitis cases and controls for putative biomarkers. A custom panel of genes was created for this expression study; the list of genes was sourced from prioritized genes based on the RegulomeDB annotation, as well as select genes in the vicinity of the lead SNPs from the GWAS.

These simultaneous analyses may enable elucidation of the specific molecular pathways involved in the development of appendicitis, and thus may open the door to improved diagnostics, treatments, and novel preventive measures for this common disease.

6.5   PUBLIC HEALTH SCREENING FOR SUSCEPTIBILITY TO ACUTE APPENDICITIS

Given the high prevalence of acute appendicitis, the emergent nature of its presentation, and the lack of specificity in current diagnostic testing, it is worthwhile to consider a personalized public health program that integrates individual genetic polymorphisms to improve the care of individuals at risk for the disease. This is especially timely: President Barack Obama recently announced a $215 million investment into the Precision Medicine Initiative, a program which aims to integrate genomic advances into clinical care and public health, and to fund further genomics research[88]. The aim of the appendicitis public health screening program would be to pre-emptively identify individuals at increased risk for appendicitis in order to improve diagnostic accuracy in case of appendicitis symptoms, and to empower individuals to invest in and manage their health. Identifying at-risk individuals would be especially beneficial given that appendicitis is most commonly a pediatric affliction, and children may not have the vocabulary nor the insight to articulate its symptoms, nor discern them from those of common stomach aches.

Appendicitis is a complex disease with genetic, bacterial, nutritional, and other environmental components, thus a thorough screening program would address multiple factors involved in its etiology.

6.5.1   Creation of a Mathematical Model for Calculation of Risk of Developing

Appendicitis and Population Health Assessment

Polymorphism data can be used in conjunction with other risk factor data to develop a

mathematical model for the likelihood of developing appendicitis at a point in time – a type of

Alvarado score for asymptomatic individuals, one that can fluctuate over the course of a patient's

life depending on the risk factors present. Genetic screening would consist of genotyping the

SNPs identified in the GWAS known to influence the risk of developing the condition (if these

SNPs are replicated in future studies and shown to be valid across populations). Because there is

no risk for neonatal presentation of appendicitis, the voluntary genetic screening portion of the

program would be instituted at an early pediatric appointment.

In a parallel effort, known risk factors for developing appendicitis would be collected on

the child and their family. First and foremost, a family history of appendicitis would be collected

and entered into the electronic medical record. If validated in further studies, risk factors that

have been reported in the literature, such as breast feeding duration, would also be collected. It

is known that the primary microbe associated with appendicitis is also associated with other

adverse health outcomes like periodontal disease, and preterm labor. Although a family history

of these conditions has not been examined with respect to risk of appendicitis to date, there exists

a plausible connection, and if validated, can be included in the family history questionnaire.

Other risk factors to include could involve a recent move of the family from a developing

country, given that this could mask the predisposition to appendicitis in previous generations and

give a falsely reassuring family history. This family history would be updated at each visit for

new factors, such as a later tonsillectomy (if its association with tonsillectomy is replicated).

Although many of the preceding risk factors listed have only been reported once in the literature,

it would not be costly to confirm their associations with the outcome of appendicitis within existing data sets. Next, principal component analysis of the genetic, family history, and environmental risk factors could be used to create a predictive model of the risk of appendicitis in the patient. The individuals would then be stratified according to their risk based on this model. Those with risk surpassing a defined threshold would be flagged for follow-up counseling regarding their risk factors and their potential mitigation. This high-risk status would also be prominently featured in the individual's electronic medical record (EMR) for consideration in the event of appendicitis-like symptoms.

A key component of the optimal use of the risk figure in emergent situations would be its integration into existing algorithms for the diagnosis of appendicitis with the aim of achieving higher diagnostic accuracy than current imaging. To ensure that emergency room physicians and surgeons are confident in the clinical benefits of using the risk figure, additional education and medical conference presentations would be provided.

6.5.2   Integration of Appendicitis Risk Data into EMRs and Patient Access

Previous translational efforts of integrating polymorphism data into clinical care have been met with certain challenges. Firstly, when genetic testing is provided at the point-of-care, there are delays in its utilization for clinical decision-making. This barrier would be unacceptable for acute appendicitis, given its rapid onset and the short timeframe for diagnosis to minimize the risk of perforation. Secondly, these translational efforts have shown that there is clinician uncertainty regarding the clinical and economic benefits of using polymorphism data to guide decision-making[89]. Both of the preceding concerns could be mitigated by preemptively genotyping the

41

patients, and providing risk analysis in advance of emergent situations, as well as providing necessary provider education.

The results of the genetic screen, family history, and final risk figure would be entered into the electronic medical record (EMR) of the patient, in accordance with the "Meaningful Use" requirement of the Health Information Exchange (HIE), the US government's initiative to allow patients and providers secure and rapid sharing of medical information electronically[90]. Should the individual in question report to the emergency department for gastrointestinal issues, the risk figure would aid the clinician in diagnosis.

The risk figure would be addressed by the primary care provider (PCP) at the high-risk patient's subsequent appointment: the patient and/or their caregiver would be educated on the symptoms of appendicitis. This risk figure would be further accessible over the internet through the patient's electronic medical record, along with resources like videos and further information on the condition. These electronic resources and the PCP conversation would promote risk-mitigating behavior change, such as recommending increased fiber intake, or promoting breast feeding on the part of the mother in future pregnancies. Knowledge of increased personal susceptibility to a condition can have a positive effect on individuals' behavior from the standpoint of prevention. Indeed, receipt of personalized dietary recommendations based on individuals' polymorphisms has been shown to effect positive long-term changes in certain nutritional intakes[91].

6.5.3   Targeted Screening Over the High-Risk Patient's Life Course: Microbial Factors

A strong association exists between the presence of *Fusobacterium nucleatum* within the appendix and acute appendicitis, but the appendix is clearly not readily accessible for direct

microbiological assessment for this risk factor. Nevertheless, it is known that microbiological disturbances in certain anatomical niches during states of disease can be reflected in microbial changes in adjacent, more readily accessible niches. For example, the diversity of the vaginal microbiome correlates with risk for preterm labor caused by infections of the reproductive tract[92], and there is evidence that the skin microbiome of individuals with diabetes has greatly increased amounts of *Staphylococcus aureaus*[93], while intestinal microbial diversity is decreased[94]. Furthermore, it was recently demonstrated that Type II Diabetes is associated with increased amounts of bacteria in serum, possibly due to translocation from the gut to the bloodstream[95]. Thus, given that there are known microbial changes associated with appendicitis, it is imperative to identify microbial biomarkers for appendicitis in more readily accessible sources, like stool or serum. If such a marker is identified, it could be assayed on a regular basis in high-risk individuals. Modern next generation sequencing methods and corresponding analytical pipelines are now capable of identifying the microbial composition of biological samples within 5 to 16 hours[96], in contrast with cell culture methods that can take several days. Interestingly, an increased amount of *F. nucleatum* in stool samples has been associated with colorectal cancer (CRC), and irritable bowel disease (IBD)[97]. Protocols for identifying and quantifying *Fusobacterium nucleatum* in stool as a sensitive and specific screen for colorectal cancer have already been developed and patented[98], and may be of use for developing screening for appendicitis within other microbial niches. Unfortunately, in patients with appendicitis, the more easily accessible microbiomes of the oral cavity and of stool do not reflect an increase in *F. nucleatum* above expected levels (Michael Morowitz, M.D., FACS, unpublished data), thus other niches may be necessary to explore for microbial screening, such as the blood.

Depending on the sensitivity and specificity of the developed screening, periodic assessment of the microbial biomarkers, or blood biomarkers of the microbial imbalance in at-risk individuals could be incorporated into the mathematical risk model for appendicitis. Regular microbial stool or serum analysis could be made more convenient to the at-risk population through the use of local sequencing stations in existing locations like pharmacies. Indeed, a private company already exists that provides rapid, inexpensive blood-based multi-analyte diagnostics through stations at local pharmacies, and it is working on pathogen detection as part of its pipeline[99,100]. Critics of technology that allows for patient self-testing such as this make the valid point that harm could result to the patient if the patient is left to interpret their own test results[100]. Therefore, it would be critical that the microbial screening results be automatically input into the patient's medical record and integrated with the global risk figure, and the physician notified of an abnormal result.

To ensure optimal integration of the screening and results into patient care, several focus groups would be held to query which geographic locations would be best for the supplementary microbial screening, and the preferred method of educating the patient and their family regarding their risk and prevention resources on a continual basis.

6.5.4 Screening Follow-Up, Surveillance, and Further Research

As outlined above, following ascertainment, the high-risk patient would be counseled regarding their risk and risk mitigation in person and using electronic methods, as well as using any other cost-effective methods identified through the focus group studies. Abnormal values on supplementary screening, and extended absence of screening would be reported to the PCP.

Population data related to risk figures, screening participation, and outcomes would be collected to surveil and evaluate to what extent the program is effective at reducing appendicitis incidence, reducing negative appendectomy rates, and reducing rates of perforated appendicitis.

A key function of public health is to conduct research to find new or improved solutions to existing public health problems. In the context of appendicitis, the initial priority would be to confirm the preliminary genetic and environmental risk factors to enable more precise elucidation of individual risk, as well as to better understand epistatic interactions among the group of associated SNPs. An additional effort would involve the integration of this risk figure into existing diagnostic algorithms used in emergency departments.

To gauge an individual's risk over time and to monitor treatment, the development of microbial or other biomarkers must be developed. Fortunately, the National Institute of Health's (NIH) Human Microbiome Project along with the European MetaHIT are involved in systematically studying the structure and function of the human gut microbiome, and this should be of use in helping to identify promising biomarkers of disease like appendicitis. However, these projects do not currently focus on identifying changes in the blood microbiome, which has been implicated in other diseases involving dysbiosis of the gut[101], thus it may be fruitful to pursue study of blood-based microbial screening methods for appendicitis.

Last but not least, an important research effort would entail designing effective preventive interventions for the high-risk population. This might include enacting novel programs to promote improved diet and increase breastfeeding rates and duration, or targeting existing evidence-based public health interventions addressing these factors to this subgroup of individuals. Additional research into interventions to mitigate the microbial component of the disease would also be warranted. For example, fecal transplantation has been used to treat other

intestinal infections such as *Clostridium difficile* with excellent results, and inflammatory conditions such as IBD[102]. It would be plausible that a similar intervention could ameliorate disturbed flora in the context of a patient at high risk for appendicitis. Additional treatments to further investigate might involve antibiotics or targeted probiotics to address the microbial component of the disease.

A broad effort to address this common disease would involve public and private stakeholders, including public health departments, governmental programs that have bearing on nutrition and determination of individual microbiomes like Women, Infants, Children (WIC), healthcare providers - especially pediatric ones, existing academic researchers, and private companies involved in developing human and microbial sequencing technologies, as well as researchers or companies involved in developing therapeutics to address disturbances of the microbiome[103].

Table 4. Putative Regulatory Elements in the Vicinity of the Four Lead SNPs.

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr3	49361791	rs13078949	1f		WDR6	
chr3	49365269	rs6795772	1f		MON1A	
					P4HTM	
					WDR6	
				CTCF		
chr3	49366741	rs9863142	1f		RNF123	
					USP4	
chr3	49370544	rs9883813	3a			FOXP3
						Nkx2-6
				CEBPB		
chr3	49378088	rs4955430	1f		APEH	
					P4HTM	
chr3	49382444	rs200101441	3a			HeliosA
						IRF3
						PU.1
						Tcfap2e
				IKZF1		
				NFATC1		
				PML		
				POLR2A		
				STAT5A		
chr3	49382614	rs113239747	2b			NFE2L2
				ATF2		
				BATF		
				BCL11A		
				CEBPB		
Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				EBF1		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				EP300		
				FOXM1		
				GATA2		
				IKZF1		
				IRF4		
				MEF2A		
				MEF2C		
				NFATC1		
				NFIC		
				NFKB1		
				PAX5		
				PML		
				POLR2A		
				POU2F2		
				RUNX3		
				SMARCA4		
				SP1		
				SPI1		
				STAT5A		
				TAF1		
				TBP		
				YY1		
chr3	49382925	rs9818758	1f		RNF123	
				CEBPB		
				GATA2		
				IKZF1		
				NFATC1		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				NFKB1		
				PML		
				POLR2A		
				SMARCA4		
				STAT5A		
				TAF1		
				TBP		
chr3	49383642	rs78574069	3a	CEBPB		
				CEBPD		
				HNF4G		
				TCF7L2		
				TEAD4		
chr3	49389143	rs56101322	2b			MEF-2
				CEBPB		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				FOS		
				FOSL2		
				SPI1		
chr3	49390250	rs17650792	1f		QARS	
					WDR6	
				POLR2A		
chr3	49393267	rs8179172	3a			E2F2
						Nanog
				POLR2A		
chr3	49394834	rs1050450	3a			AP-4
				POLR2A		
chr3	49396360	rs3811699	2b			MZF1
chr3	49396751	rs3448	1b			RP58
					APEH	
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
					P4HTM	
				ERG		
chr3	49397284	rs8179164	2b			GATA-2
				CTCF		
				FOS		
				FOSL2		
				GATA1		
				JUN		
				JUND		
chr3	49411404	rs7621003	1f		QARS	
					WDR6	
chr3	49423976	rs6797765	1a		QARS	
					WDR6	
				CEBPB		
chr3	49425180	rs138478251	2b			Elf3
						Foxd3
						FOXP1
						Srf
						Tcfap2e
						Zfp105
						FOXC1
						FOXD3
				FOS		
chr3	49439440	rs2140270	1f		APEH	
					P4HTM	
chr3	49439725	rs2177268	1f			FOXP1
					AMT	

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr3	49443081	rs13096474	1f		AMT	
					APEH	
					NCKIPSD	
					NICN1	
					P4HTM	
					WDR6	
chr3	49448583	rs79186983	2b			FOXJ3
						FOXO1
						FOXP1
						FOXK1
						Srf
				POLR2A		
chr3	49448659	rs4855875	3a			FIGLA
				POLR2A		
				RBBP5		
				SIN3A		
chr3	49448785	rs4855874	3a			FOXG1
				HNF4A		
				POLR2A		
				RBBP5		
				SIN3A		
				UBTF		
chr3	49448818	rs6777731	3a			Sp100
				E2F1		
				HNF4A		
				POLR2A		
				RBBP5		
				SIN3A		
				UBTF		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
chr3	49449638	rs940045	1f		APEH	
					P4HTM	
chr3	49449685	rs139004176	2b			NRF1
chr3	49450864	rs6784820	1f		AMT	
					NICN1	
					QARS	
					WDR6	
				CTCF		
				MAX		
				MYC		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				RAD21		
				USF1		
chr3	49453834	rs6997	1f		USP4	
				EBF1		
				ELF1		
				FOXP2		
				MYC		
				NFATC1		
				NFIC		
				PML		
				POLR2A		
				RUNX3		
				SIN3A		
				SPI1		
chr3	49454112	rs9814873	1f		USP4	
				NFIC		
				EBF1		
chr3	49455330	rs11715915	1f		USP4	
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				POLR2A		
					QARS	
					WDR6	
chr3	49459114	rs4855873	1d		QARS	
					WDR6	
				NFIC		
				PML		
chr3	49459252	rs1464567	1f		AMT	
					NICN1	
					USP19	
					WDR6	
				PML		
				POLR2A		
chr3	49459376	rs1464566	1b		QARS	
					WDR6	
				FOXA2		
				POLR2A		
chr3	49460350	rs1464569	1f		APEH	
					P4HTM	
				EP300		
				ESRRA		
				FOXA1		
				FOXA2		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				HDAC2		
				HNF4A		
				MAX		
				MXI1		
				MYBL2		
				POLR2A		
				RXRA		
				SP1		
				TBP		
				TEAD4		
chr3	49460407	rs8897	1d		AMT	
					APEH	
					NCKIPSD	
					NICN1	
					P4HTM	
					WDR6	
				ESRRA		
				FOXA1		
				FOXA2		
				HNF4A		
				MAX		
				MXI1		
				MYBL2		
				POLR2A		
				RXRA		
				TBP		
				TEAD4		
chr3	49465162	rs73088161	3a			AR
				FOS		
						Zbtb3
				FOXA1		
				FOXA2		
chr3	49497743	rs76711745	3a	TFAP2A		
Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr3	49497883	rs885592	1d			Zbtb3
					QARS	
					WDR6	
				TFAP2A		
				TFAP2C		
chr3	49499167	rs11130197	2b			ZNF75A

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				CHD1		
				CTCF		
				EP300		
				HNF4A		
				RAD21		
				RBBP5		
				SIN3A		
				SP1		
				TAF1		
				TBP		
				TCF12		
				TEAD4		
				TFAP2C		
				YY1		
chr3	49499829	rs10865955	1f		QARS	
					WDR6	
				AR		
				ATF1		
				EP300		
				FOSL1		
				FOXA1		
				JUN		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				JUND		
chr3	49502504	rs139173102	2b	CEBPB		
				FOS		
chr3	49502548	rs149926066	3a			aMEF-2
						Cdx
						FOXC1
						FOXC2
						Foxl1
						FOXP1
						HNF3alpha
						HNF3beta
						MEF-2
						ONECUT3
						Elf3
						Srf
				CEBPB		
				FOS		
chr3	49503625	rs80267491	3a			ESRRB
				FOXA2		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				GATA1		
				HNF4A		
				POLR2A		
				TFAP2A		
chr3	49506900	rs9860055	1f		APEH	
					P4HTM	
				MTA3		
				POLR2A		
				RFX3		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				SIN3A		
				SMARCB1		
chr3	49507030	rs150568866	2b			Zfp187
						ZNF524
chr3	49507668	rs187508379	2b			CNOT3
chr3	49508971	rs11456203	2b	TAL1		
chr3	49508976	rs7613491	2b	TAL1		
chr3	49521974	rs4855864	1f			E2F3
					QARS	
					WDR6	
chr3	49522822	rs115890970	2b			HOXD3
						IPF1
				BHLHE40		
				EP300		
				NFIC		
chr3	49535115	rs7637999	1f		QARS	
					WDR6	
chr3	49538799	rs11130199	1f		QARS	
					WDR6	
				CEBPB		
				EP300		
				FOS		
				GATA1		
				JUND		
				MYBL2		
				POLR2A		
				TAL1		
chr3	49540389	rs79873387	3a	FOS		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr3	49546864	rs78407378	3a			BCL6
				FOXA2		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				HNF4A		
chr3	49550481	rs186347793	3a			Foxj3
				GATA6		
chr3	49555583	rs13322887	1f	APEH		
				P4HTM		
chr3	49557051	rs3870338	1f	AMT		
				NICN1		
				QARS		
				WDR6		
chr3	49557857	rs3870336	1f	ARIH2		
				DAG1		
				RBM6		
chr3	49570882	rs1050088	1f	AMT		
				NICN1		
				QARS		
				WDR6		
chr3	49571462	rs12583	1f	APEH		
				P4HTM		
chr3	49572140	rs4625	2b			Bach1
				CEBPB		
chr3	49572403	rs11538155	2b			MAZR
						Zfp281
				CEBPB		
chr3	49572894	rs6446283	3a			HNF4A
				EP300		
				MAX		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				MXI1		
				SIN3A		
				YY1		
chr3	49574808	rs13079082	3a			LUN-1
				CEBPB		
chr3	49577592	rs74426367	2c			Tcfap2b
						Tfap2c
chr3	49577665	rs1982861	1f	APEH		
				P4HTM		
chr3	49577960	rs78040846	2b			Ikaros
						Sox13
chr3	49581559	rs10865956	1f	APEH		
				P4HTM		
chr3	49591539	rs115713947	3a			Irf4
chr3	49598064	rs9862534	1f	APEH		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
					P4HTM	
chr3	49600319	rs4241407	1f		AMT	
					APEH	
					NCKIPSD	
					NICN1	
					P4HTM	
					WDR6	
				AR		
chr3	49600426	rs4241406	2b			Mrg2
						TFAP4
						TGIF
						TGIF1
						Tgif2

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				AR		
chr3	49601255	rs2312462	1f		AMT	
					APEH	
					NCKIPSD	
					NICN1	
					P4HTM	
					WDR6	
chr3	49625093	rs72936019	2a			FOXG1
						NF-Y
						NFYA
chr3	49639803	rs1491983	1d			AP-4
						Lmo2complex
					RHOA	
				ESR1		
				TFAP2A		
chr3	49645209	rs11130211	1f		HEMK	
chr3	49646981	rs2029591	1f		AMT	
					NICN1	
					RHOA	
					WDR6	
chr3	49649434	rs9823134	2b	FOS		
chr3	49655927	rs62262672	3a			FOXP3
				REST		
chr3	49657441	rs4855833	1f		RHOA	
chr3	49658084	rs13096480	1f		COX4NB	
				EGR1		
chr3	49665390	rs2131108	1f		AMT	
					NICN1	

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
					RHOA	
chr3	49667691	rs9883000	1f		NICN1	
					RHOA	
chr3	49674343	rs9824435	1f		NICN1	
					RHOA	
				FOS		
chr3	49679072	rs2329021	1f		RHOA	
chr3	49685073	rs2329020	1f		WDR6	
chr3	49685592	rs1078341	1b			HMX2
						HMX3
						Nkx2-4
						Nkx2-6
					NICN1	
					RHOA	
				MYC		
				USF1		
chr3	49687779	rs6774202	1b			HP1sitefactor
					RHOA	
				BATF		
				NFKB1		
				RUNX3		
				ZNF263		
chr3	49690199	rs4855885	1f		RHOA	
				ZNF263		
chr3	49696633	rs2131104	1f		WDR6	
chr3	49696797	rs11718165	1f		C3orf62	
					DAG1	
					HSS00328072	
Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
					MST1	
					USP4	
chr3	49701298	rs2005557	2b			NRSF
						REST
				MYC		
chr3	49701983	rs9858542	1f		USP4	
chr3	49706403	rs187773800	2b			Zfp410
						Zfp281
						RREB1
						CKROX
						MAZ

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
						SP1
chr3	49708502	rs1060962	2b	MXI1		
chr3	49708590	rs34560231	3a	NFIC		
chr3	49708769	rs1060970	1b			EWSR1-FLI1
					AMIGO3	
					GMPPB	
					UBA7	
				EP300		
				FOXA1		
				FOXA2		
				HNF4A		
				MYBL2		
				NFIC		
				RXRA		
				SP1		
chr3	49708807	rs35637631	2b	EP300		
				FOXA1		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				FOXA2		
				HNF4A		
				MYBL2		
				NFIC		
				RXRA		
				SP1		
				USF1		
chr3	49709147	rs71324984	2b			Hbp1
						Nanog
chr3	49711430	rs140606103	2b			VDR=CAR=PXR
chr3	49711559	rs113530503	2b			AP-2
chr3	49712220	rs111635853	2b			SP1
						WT1
						Zfp281
chr3	49715446	rs4855881	1f		WDR6	
chr3	49719729	rs9822268	3a			SP1
				POLR2A		
chr3	49720044	rs34491127	2b			RFX1
				POLR2A		
chr3	49720887	rs73834009	3a			EGR1
						EGR2
						EGR3
						EGR4
						Zif268

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				NR3C1		
				POLR2A		
chr3	49723001	rs201720919	3a	POLR2A		
chr3	49724534	rs41291700	2b			MAF
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				POLR2A		
chr3	49725859	rs145408661	2b			PLAG1
				NFIC		
				POLR2A		
				TAF1		
chr3	49726555	rs4052565	2b			CACD
chr3	49727754	rs140793082	3a			RFX5
				GATA1		
				REST		
				TAL1		
chr3	49727886	rs7373192	2b	REST		
				TAL1		
chr3	49735746	rs11130214	1f			SPDEF
					WDR6	
chr3	49736269	rs79587292	3a	GATA1		
chr3	49737323	rs11130217	1f		RHOA	
chr3	49742107	rs34154145	2b	POLR2A		
chr3	49745235	rs11709734	1f		RHOA	
chr3	49750261	rs77208503	2b	SPI1		
chr3	49751585	rs2291542	1f		WDR6	
chr3	49751856	rs12715437	1f		RHOA	
chr3	49753003	rs34614773	3a			ZIC3
						ZIC4
chr3	49753788	rs11720705	1f		RHOA	
				POLR2A		
chr3	49753901	rs61743872	3a			CAC-bindingprotein
				POLR2A		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
chr3	49754970	rs7628207	1f		RHOA	
				POLR2A		
chr3	49755676	rs62640368	2b			CACCC-bindingfactor
				POLR2A		
chr3	49756212	rs35201844	2b			Ascl2
						E2A
				POLR2A		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr3	49758111	rs7634945	2b			ER
						ESR1
				POLR2A		
chr3	49758497	rs34127462	3a			Bcl6b
				POLR2A		
				MTA3		
chr3	49758764	rs4768	2b			YY1
				MTA3		
				POLR2A		
chr3	49760431	rs34345884	2b			NRSF
						REST
				POLR2A		
chr3	49760477	rs11547261	2b			CREB5
						JDP2
						Pax-3
				POLR2A		
chr3	49761613	rs3811695	1f		(dsQTL)	
chr3	49771990	rs9849038	1f		RHOA	
chr3	49808981	rs11717463	3a			FOXP1
				MEF2A		
				RUNX3		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
chr3	49813258	rs9853352	1f		RHOA	
					USP4	
					WDR6	
chr3	49817450	rs9829155	1f		RHOA	
chr3	49818555	rs9814765	2c	CEBPB		
chr3	49823200	rs60844685	3a	E2F1		
				GATA1		
				MAX		
				SPI1		
chr3	49827791	rs55750059	2b			Six-1
						Six-6
chr3	49828863	rs145439991	3a			Egr1
				JUND		
chr3	49829653	rs7637711	1f		NICN1	
					RHOA	
chr3	49832788	rs73079003	3a			KLF16
						Klf7
						Sp1
						SP3
				HNF4A		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr3	49834571	rs6809879	1f		RHOA	
chr3	49840525	rs201878414	2b			NFKB1
chr3	49840843	rs115355405	3a			Meis1
						Mrg2
						Pknox1
						MEIS2
						Tgif2
chr3	49841310	rs78926068	3a			NFKB1
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
chr3	49842881	rs11400401	2b			mTERF
						RREB-1
				SMARCA4		
				ETS1		
				MYC		
				POLR2A		
				RAD21		
				GABPA		
chr3	49843723	rs3819325	1b			CAC-bindingprotein
						Egr
						RREB1
						SP1
						SP4
						Zfp281
						Zfp740
						ZNF219
						ZNF515
						ZNF740
				CHD2		
				ETS1		
				MYC		
chr3	49844001	rs72938113	3a			TFAP2C
				CHD2		
				EP300		
				ETS1		
				GABPA		
				GATA1		
				GATA2		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				MAZ		
				MYC		
				POLR2A		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				RCOR1		
				SMARCA4		
				TAL1		
				TEAD4		
chr3	49845006	rs58339610	3a			Pknox1
						PKNOX2
				ELF1		
				EP300		
				NFKB1		
				PBX3		
				POLR2A		
				SMARCA4		
				TCF12		
				TCF3		
chr3	49860854	rs6446298	1f			CART1
					AK097846	
					APEH	
					RNF123	
				MAFK		
chr3	49878078	rs2271960	1f		RBM6	
				CTCF		
chr3	49878113	rs2271961	1f		RBM6	
				CTCF		
chr3	49878264	rs1996663	1f		AMT	
					NICN1	
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
					RHOA	
				CTCF		
chr3	49878395	rs1996664	1f		RHOA	
				CTCF		
chr3	49878779	rs75160702	3a			FIGLA
chr3	49880399	rs34484573	2b	USF1		
chr3	49880943	rs1110295	1f			CAC-bindingprotein
					HEMK	
				TCF7L2		
chr3	49894030	rs2276864	1f		-	
chr3	49896727	rs111272205	2b			MAZR
						PPARalpha:RXRalpha
						SP1
						UF1H3BETA
				CTCF		
chr3	49901060	rs11130222	2b			LUN-1

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
						NHLH1
				TEAD4		
chr3	49902160	rs2883059	1d		HYAL3	
					RBM6	
				MAFF		
				MAFK		
chr3	49911354	rs77491796	2b			NKX2-3
						Nkx2-5
						T
						TBX20
						TRUE
				CDX2		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				CTCF		
				HNF4A		
				MXI1		
				YY1		
chr3	49915506	rs9862795	2b			DMRT7
				POLR2A		
chr3	49918751	rs9813644	1b		MON1A	
				RFX3		
chr3	49918975	rs58943948	2b	POLR2A		
				RFX3		
chr3	49920297	rs56352827	2b	POLR2A		
chr3	49924328	rs7615240	2b			AIRE
				POLR2A		
chr3	49936102	rs2230590	2b	MYC		
chr3	49936715	rs41291716	2b			ESRRA
						ESRRG
						EWSR1-FLI1
						RARA
						RARB
						RARG
				BHLHE40		
				CTCF		
				RAD21		
				POLR2A		
chr3	49943570	rs75195683	2b			Egr
						KROX
chr3	49945996	rs58025924	2b			LRF
						Plagl1
Chromosome	**Coordinate**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**

Chromosome	Coordinate	rsID	Score	Protein	eQTL	Motif
	(1-based)					
						ZIC1
						ZIC3
						ZIC4
				GATA1		
				MAZ		
chr3	49948627	rs201449228	2b			Evi-1
				GATA1		
				GATA2		
				POU5F1		
				RBBP5		
chr3	49977786	rs116486986	2b			RNF96
chr3	49997963	rs2624843	3a	GATA1		
chr3	49998282	rs2883057	3a	GATA1		
chr3	50004209	rs2252833	1d		RBM6	
				FOS		
				RFX3		
chr3	50028246	rs2248256	3a			TP53
						TP63
						TP73
				YY1		
chr3	50039474	rs7628058	1f		RBM6	
chr3	50044006	rs7635601	1f		HYAL3	
					RBM6	
chr3	50082914	rs2526747	1f		HYAL3	
					RBM6	
chr3	50087947	rs144406288	2b	EP300		
				FOS		
				FOXA1		
Chromosome	**Coordinate**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
	(1-based)					
				MAFF		
				MAFK		
				POLR2A		
chr3	50098337	rs17050913	1f		HEMK	
chr15	73259907	rs1449270	3a			Eomes
						TBR1
						TBX2
				TRIM28		
chr15	73260071	rs28440854	3a	TRIM28		
chr15	73260649	rs28666516	3a	TCF7L2		
				TRIM28		
chr15	73263049	rs8025665	1f			IRF8

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
					KCNS1	
chr15	73275132	rs79403000	3a	IKZF1		
chr15	73275175	rs8033860	3a	IKZF1		
chr15	73275178	rs8034120	3a			PPARG::RXRA
				IKZF1		
chr15	73276287	rs4522396	3a	FOS		
chr15	73276395	rs4260030	3a			TTF1(Nkx2-1)
				FOS		
chr15	73283449	rs75940106	3a			Barhl2
						FOXP1
						HMX2
						Hmx3
						Hoxb13
						ISX
						LHX9
						MSX1
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
						MSX2
						Msx3
						Nkx5-2
						Nkx6-1
						PRRX2
						SHOX2
				FOS		
				MYC		
chr15	73284093	rs11072401	2b			MEF-2
						MEF2A
						MEF2B
						MEF2D
				STAT3		
chr15	73284181	rs9920504	1b			Elf-1
					NEO1	
				FOS		
				STAT3		
chr15	73284535	rs9920548	3a	CEBPB		
				FOS		
chr15	73287152	rs7172316	2b	TCF7L2		
chr15	73287611	rs78572431	3a			HSF1
				TCF7L2		
chr15	73290627	rs9920770	3a			Roaz
						TTF-1(Nkx2-1)
				MYC		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				STAT3		
chr15	73290692	rs7496513	3a	MYC		
				STAT3		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr15	73290939	rs138108225	3a	STAT3		
chr15	73305445	rs79993429	3a			Gata5
				IKZF1		
chr15	73313006	rs28386776	3a	HNF4A		
chr15	73319108	rs76778349	2b			CACD
						KLF16
						SP3
						Spz1
						UF1H3BETA
						Zfp281
						ZNF219
				MAX		
chr15	73323136	rs112949915	2b			PPARalpha:RXRalpha
				RFX3		
chr15	73327983	rs78424364	3a	GATA1		
chr15	73343331	rs55969660	3a			Oct-4(POU5F1)
				EZH2		
chr15	73343578	rs141982291	3a			ZBTB7B
				EZH2		
chr15	73343619	rs77667665	3a	EZH2		
chr15	73344196	rs62016793	2b			IRF-2
						IRF8
chr15	73347229	rs116353503	3a			FOXP1
				AR		
				IKZF1		
chr15	73371072	rs116463863	3a			EWSR1-FLI1
				GATA3		
chr15	73371315	rs68000913	2b			core-bindingfactor

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr15	73384345	rs8033192	3a	GATA1		
chr15	73384481	rs77918095	3a	GATA1		
chr15	73406631	rs73440033	3a	USF1		
chr15	73417826	rs8026579	3a	CEBPB		
chr15	73419767	rs9972347	2b			RREB1
						ZNF784
chr15	73424712	rs80200465	3a	IKZF1		
chr15	73432078	rs115253960	3a			Hoxa13

Chromosome	Coordinate	rsID	Score	Protein	eQTL	Motif
						HOXC13
				GATA2		
				GATA3		
chr15	73443374	rs147723618	3a			ZNF75A
				CTCF		
				SPDEF		
chr15	73466694	rs11636981	3a			TBX15
				NFKB1		
chr15	73472096	rs74025262	3b			AR
				AR		
chr15	73472104	rs139043297	3b			TEF
				AR		
chr15	73513938	rs12373012	3a	POLR2A		
chr15	73513993	rs118016492	3a			UF1H3BETA
				POLR2A		
chr15	73524629	rs2046017	3a			FOXJ3
				MAFK		
chr15	73530356	rs147551053	2b			NRSE
				BACH1		
				EP300		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				GATA2		
				GATA3		
chr15	73530599	rs192120450	2b			IRF1
						Isgf3g
				BACH1		
				EP300		
chr15	73532372	rs1714530	3a			Elk-1
				CTCF		
				RAD21		
chr15	73557903	rs2660825	2b			MRF-2
				CTCF		
				CTCF		
				MYC		
				MYC		
				NFATC1		
				SIN3A		
chr15	73565650	rs8027588	2b			PRDM1
						STAT1
				CTCF		
chr15	73576507	rs115719976	2b			Evi-1
						Srf

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				CTCF		
				RAD21		
chr15	73586171	rs41415044	2b			Mtf1
						NF-AT
				HDAC2		
				HNF4A		
				SP1		

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
				HNF4A		
				EP300		
				HDAC2		
				SP1		
chr15	73586267	rs2252725	3a			Zbtb3
				EP300		
				FOXA1		
				FOXA2		
				HDAC2		
				HNF4A		
				HNF4G		
				JUND		
				SP1		
chr15	73586515	rs62015483	3a			RFX1(EF-C)
						RFX2
						RFX3
						RFX4
				HNF4A		
chr15	73586577	rs148534572	3a	HNF4A		
chr15	73588382	rs576813	2b			LRF
				CEBPB		
				FOS		
				MAX		
				MYC		
chr15	73590153	rs74022931	3a			Nkx2-5
				IKZF1		
chr15	73598607	rs531019	2b			FOXfactors
						HNF3

Chromosome	Coordinate (1-based)	rsID	Score	Protein	eQTL	Motif
chr15	73605644	rs11630633	3a			FOXP1
				BACH1		
				MAFF		
				MAFK		

Chromosome	Coordinate	rsID	Score	Protein	eQTL	Motif
chr15	73619997	rs494493	2b			RARB
				IKZF1		
chr15	73623689	rs572112	3a	IKZF1		
chr15	73623691	rs2680333	3a	IKZF1		
chr15	73624849	rs76789737	2b			p53
						TP53
				IKZF1		
chr15	73632376	rs4776632	3a			Bcl6b
				CTCF		
chr4	111656967	rs11944778	3a			ONECUT1
						ONECUT3
						Oct-1
						FOXP1
				ZNF263		
chr4	111663478	rs7439625	3a	FOS		
chr4	111677661	rs74843677	3a	CTCF		
				RAD21		
chr4	111708438	rs7434417	3a	CEBPB		
				FOS		
						BARX1
chr4	111714889	rs2220427	2b	GATA6		
				HNF4A		
chr4	111719978	rs17042198	3a	FOXP2		
chr4	112769004	rs385040	3a	CTCF		
Chromosome	**Coordinate (1-based)**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
				RAD21		
chr4	112843090	rs72899903	3a	FOS		
chr4	112845202	rs76743823	2b			Foxl1
						FOXJ2
						FOXJ3
				CTCF		
				FOS		
chr4	112846378	rs17589101	3a			REST
				CTCF		
chr4	112858746	rs7670652	2b			Dobox4
						Evi-1
chr4	112865123	rs72899924	2b			GCM
chr4	112873874	rs115874059	3a			FOXO3A
				POLR2A		
chr4	112876056	rs75690152	3a			FOXJ2
				CEBPB		
chr4	112881725	rs886567	3a			NGFI-C

					Egr-2
					EGR1
					Zif268
				GATA3	
				POLR2A	
				TFAP2A	
				GATA1	
				SMARCA4	
chr4	112883200	rs757507	3a	EP300	
				FOS	
					Bsx

SNPs and their corresponding genetic elements with a RegulomeDB score of less than or equal to 3 are displayed.

Table 5. Motifs Identified within the Region of Interest on Chromosome 3 and their Targets.

Displayed are the position weight matrices (PWMs) of the motifs organized by putative regulatory variant. The number of entries of the motif reflects the quantity of sources that identified this motif, if greater than one. The red box outlines the location of the putative regulatory variant. RegulomeDB was used to retrieve these motifs. *Table 5 attachment (.xls)*

Table 6. Motifs Identified within the Region of Interest on Chromosome 15 and their Targets.

Displayed are the position weight matrices (PWMs) of the motifs organized by putative regulatory variant. The number of entries of the motif reflects the quantity of sources that identified this motif, if greater than one. The red box outlines the location of the putative regulatory variant. RegulomeDB was used to retrieve these motifs. *Table 6 attachment (.xls)*

Table 7.Motifs Identified within the Region of Interest on Chromosome 4 (at 111.7 Mb) and their Targets.

Displayed are the position weight matrices (PWMs) of the motifs organized by putative regulatory variant. The number of entries of the motif reflects the quantity of sources that identified this motif, if greater than one. The red box outlines the location of the putative regulatory variant. RegulomeDB was used to retrieve these motifs. *Table 7 attachment (.xls)*

Table 8. Motifs Identified within the Region of Interest on Chromosome 4 (at 112.7 Mb) and their Targets.

Displayed are the position weight matrices (PWMs) of the motifs organized by putative regulatory variant. The number of entries of the motif reflects the quantity of sources that identified this motif, if greater than one. The red box outlines the location of the putative regulatory variant. RegulomeDB was used to retrieve these motifs. *Table 8 attachment (.xls)*

Table 9. Putative Regulatory Elements Organized by Distance of the Element's Modifying Variant from Lead SNP on Chromosome 3.

Chr	Coordinate	Distance (bp): within 10 kb	rsID	Score	Protein	eQTL	Motif
3	49880943	1406	rs1110295	1f			CAC-bindingprotein
				1f		HEMK	
				1f	TCF7L2		
3	49880399	1950	rs34484573	2b	USF1		
3	49878779	3570	rs75160702	3a			FIGLA
3	49878395	3954	rs1996664	1f		RHOA	
				1f	CTCF		
3	49878264	4085	rs1996663	1f		AMT	
				1f		NICN1	
				1f		RHOA	
				1f	CTCF		
3	49878113	4236	rs2271961	1f		RBM6	
				1f	CTCF		
3	49878078	4271	rs2271960	1f		RBM6	
				1f	CTCF		
Chr	**Coordinate**	**Distance (bp): within 50 kb**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
3	49894030	11681	rs2276864	1f		(dsQTL)	
3	49896727	14378	rs111272205	2b			MAZR
							PPARalpha:RXRalpha
							SP1
							UF1H3BETA
					CTCF		
3	49901060	18711	rs11130222	2b			LUN-1
							NHLH1
					TEAD4		
3	49902160	19811	rs2883059	1d		HYAL3	
						RBM6	
					MAFF		
					MAFK		
3	49860854	21495	rs6446298	1f			CART1
						AK097846	
						APEH	
						RNF123	
					MAFK		
3	49911354	29005	rs77491796	2b			NKX2-3
							Nkx2-5

							T
							TBX20
							TRUE
					CDX2		
					CTCF		
					HNF4A		
					MXI1		
					YY1		
3	49915506	33157	rs9862795	2b			DMRT7
					POLR2A		
3	49918751	36402	rs9813644	1b		MON1A	
					RFX3		
3	49918975	36626	rs58943948	2b	POLR2A		
					RFX3		
3	49845006	37343	rs58339610	3a			Pknox1
							PKNOX2
					ELF1		
					EP300		
					NFKB1		
					PBX3		
					POLR2A		
					SMARCA4		
					TCF12		
					TCF3		
3	49920297	37948	rs56352827	2b	POLR2A		
3	49844001	38348	rs72938113	3a			TFAP2C
					CHD2		
					EP300		
					ETS1		
					GABPA		
					GATA1		
					GATA2		
					MAZ		
					MYC		
					POLR2A		
					RCOR1		
					SMARCA4		
					TAL1		
					TEAD4		
3	49843723	38626	rs3819325	1b			CAC-bindingprotein
							Egr
							RREB1
							SP1

Chr	Coordinate	Distance (bp): within 100 kb	rsID	Score	Protein	eQTL	Motif
							SP4
							Zfp281
							Zfp740
							ZNF219
							ZNF515
							ZNF740
					CHD2		
					ETS1		
					MYC		
3	49842881	39468	rs11400401	2b			mTERF
							RREB-1
					SMARCA4		
					ETS1		
					MYC		
					POLR2A		
					RAD21		
					GABPA		
3	49841310	41039	rs78926068	3a			NFKB1
3	49840843	41506	rs115355405	3a			Meis1
							Mrg2
							Pknox1
							MEIS2
							Tgif2
3	49840525	41824	rs201878414	2b			NFKB1
3	49924328	41979	rs7615240	2b			AIRE
					POLR2A		
3	49834571	47778	rs6809879	1f		RHOA	
3	49832788	49561	rs73079003	3a			KLF16
							Klf7
							Sp1
							SP3
					HNF4A		
Chr	Coordinate	Distance (bp): within 100 kb	rsID	Score	Protein	eQTL	Motif
3	49829653	52696	rs7637711	1f		NICN1	
						RHOA	
3	49828863	53486	rs145439991	3a			Egr1
					JUND		
3	49936102	53753	rs2230590	2b	MYC		
3	49936715	54366	rs41291716	2b			ESRRA
							ESRRG
							EWSR1-FLI1
							RARA

Chr	Coordinate	Distance (bp): within 150 kb	rsID	Score	Protein	eQTL	Motif
							RARB
							RARG
					BHLHE40		
					CTCF		
					RAD21		
3	49827791	54558	rs55750059	2b			Six-1
							Six-6
3	49938758	56409	rs7616171	2b	POLR2A		
3	49823200	59149	rs60844685	3a	E2F1		
					GATA1		
					MAX		
					SPI1		
3	49943570	61221	rs75195683	2b			Egr
							KROX
3	49945996	63647	rs58025924	2b			LRF
							Plagl1
							ZIC1
							ZIC3
							ZIC4
					GATA1		
					MAZ		
3	49818555	63794	rs9814765	2c	CEBPB		
3	49817450	64899	rs9829155	1f		RHOA	
3	49948627	66278	rs201449228	2b			Evi-1
					GATA1		
					GATA2		
					POU5F1		
					RBBP5		
3	49813258	69091	rs9853352	1f		RHOA	
						USP4	
						WDR6	
3	49808981	73368	rs11717463	3a			FOXP1
					MEF2A		
					RUNX3		
3	49977786	95437	rs116486986	2b			RNF96
Chr	**Coordinate**	**Distance (bp): within 150 kb**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
3	49771990	110359	rs9849038	1f		RHOA	
3	49997963	115614	rs2624843	3a	GATA1		
3	49998282	115933	rs2883057	3a	GATA1		
3	49761613	120736	rs3811695	1f		(dsQTL)	
3	50004209	121860	rs2252833	1d		RBM6	
					FOS		

Chr	Position	Number	rsID	Cat	Gene1	Gene2	Gene3
							RFX3
3	49760477	121872	rs11547261	2b			CREB5
							JDP2
							Pax-3
					POLR2A		
3	49760431	121918	rs34345884	2b			NRSF
							REST
					POLR2A		
3	49758764	123585	rs4768	2b			YY1
					MTA3		
					POLR2A		
3	49758497	123852	rs34127462	3a			Bcl6b
					POLR2A		
					MTA3		
3	49758111	124238	rs7634945	2b			ER
							ESR1
					POLR2A		
3	49756212	126137	rs35201844	2b			Ascl2
							E2A
					POLR2A		
3	49755676	126673	rs62640368	2b			CACCC-bindingfactor
					POLR2A		
3	49754970	127379	rs7628207	1f		RHOA	
					POLR2A		
3	49753901	128448	rs61743872	3a			CAC-bindingprotein
					POLR2A		
3	49753788	128561	rs11720705	1f		RHOA	
					POLR2A		
3	49753003	129346	rs34614773	3a			ZIC3
							ZIC4
					POLR2A		
3	49751856	130493	rs12715437	1f		RHOA	
3	49751585	130764	rs2291542	1f		WDR6	
3	49750261	132088	rs77208503	2b	SPI1		
3	49745235	137114	rs11709734	1f		RHOA	
3	49742107	140242	rs34154145	2b	POLR2A		
3	49737323	145026	rs11130217	1f		RHOA	
3	50028246	145897	rs2248256	3a			TP53
							TP63
							TP73
					YY1		
3	49736269	146080	rs79587292	3a	GATA1		
3	49735746	146603	rs11130214	1f			SPDEF

							WDR6

The coordinate of the lead SNP (rsID rs2247036) is 49,882,349. Coordinates are 1-based. Chr stands for chromosome.

Table 10. Putative Regulatory Elements Organized by Distance of the Element's Modifying Variant from Lead SNP on Chromosome 15.

Chr	Coordinate	Distance (bp):	rsID	Score	Protein	eQTL	Motif
		within 10 kb					
15	73598607	1363	rs531019	2b			FOXfactors
							HNF3
15	73605644	5674	rs11630633	3a			FOXP1
					BACH1		
					MAFF		
					MAFK		
15	73590153	9817	rs74022931	3a			Nkx2-5
					IKZF1		
Chr	Coordinate	Distance (bp):	rsID	Score	Protein	eQTL	Motif
		within 50 kb					
15	73588382	11588	rs576813	2b			LRF
					CEBPB		
					FOS		
					MAX		
					MYC		
15	73586577	13393	rs148534572	3a	HNF4A		
15	73586515	13455	rs62015483	3a			RFX1(EF-C)
							RFX2
							RFX3
							RFX4
					HNF4A		
15	73586267	13703	rs2252725	3a			Zbtb3
					EP300		
					FOXA1		
					FOXA2		
					HDAC2		
					HNF4A		
					HNF4G		
					JUND		
					SP1		
15	73586171	13799	rs41415044	2b			Mtf1
							NF-AT

Chr	Coordinate	Distance (bp): within 100 kb	rsID	Score	Protein	eQTL	Motif
					HDAC2		
					HNF4A		
					SP1		
					HNF4A		
					EP300		
					HDAC2		
					SP1		
15	73619997	20027	rs494493	2b			RARB
					IKZF1		
15	73576507	23463	rs115719976	2b			Evi-1
							Srf
					CTCF		
					RAD21		
15	73623689	23719	rs572112	3a	IKZF1		
15	73623691	23721	rs2680333	3a	IKZF1		
15	73624849	24879	rs76789737	2b			p53
							TP53
					IKZF1		
15	73632376	32406	rs4776632	3a			Bcl6b
					CTCF		
15	73565650	34320	rs8027588	2b			PRDM1
							STAT1
					CTCF		
15	73557903	42067	rs2660825	2b			MRF-2
					CTCF		
					CTCF		
					MYC		
					MYC		
					NFATC1		
					SIN3A		
Chr	**Coordinate**	**Distance (bp): within 100 kb**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
15	73532372	67598	rs1714530	3a			Elk-1
					CTCF		
					RAD21		
15	73530599	69371	rs192120450	2b			IRF1
							Isgf3g
					BACH1		
					EP300		
15	73530356	69614	rs147551053	2b			NRSE
					BACH1		
					EP300		
					GATA2		

Chr	Coordinate	Distance (bp): within 150 kb	rsID	Score	Protein	eQTL	Motif
					GATA3		
15	73524629	75341	rs2046017	3a			FOXJ3
					MAFK		
15	73513993	85977	rs118016492	3a			UF1H3BETA
					POLR2A		
15	73513938	86032	rs12373012	3a	POLR2A		
Chr	**Coordinate**	**Distance (bp): within 150 kb**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
15	73472104	127866	rs139043297	3b			TEF
					AR		
15	73472096	127874	rs74025262	3b			AR
					AR		
15	73466694	133276	rs11636981	3a			TBX15
					NFKB1		

The coordinate of the lead SNP (rs192656182) is 73,599,970. Coordinates are 1-based. Chr stands for chromosome.

Table 11. Putative Regulatory Elements Organized by Distance of the Element's Modifying Variant from Lead SNP on Chromosome 4 (at 111.7 Mb).

Chr	Coordinate	Distance (bp): within 10 kb	rsID	Score	Protein	eQTL	Motif
4	111719978	1019	rs17042198	3a	FOXP2		
4	111714889	6108	rs2220427	2b	GATA6		
					HNF4A		
Chr	**Coordinate**	**Distance (bp): within 50 kb**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
4	111708438	12559	rs7434417	3a	CEBPB		
					FOS		
							BARX1
4	111677661	43336	rs74843677	3a	CTCF		
					RAD21		
Chr	**Coordinate**	**Distance (bp): within 100 kb**	**rsID**	**Score**	**Protein**	**eQTL**	**Motif**
4	111663478	57519	rs7439625	3a	FOS		
4	111656967	64030	rs11944778	3a			ONECUT1
							ONECUT3
							Oct-1
							FOXP1
					ZNF263		

78

The coordinate of the lead SNP (rs2129979) is 111,720,997. Coordinates are 1-based. Chr stands for chromosome.

Table 12. Putative Regulatory Elements Organized by Distance of the Element's Modifying Variant from Lead SNP on Chromosome 4 (at 112.7 Mb).

Chr	Coordinate	Distance (bp):	rsID	Score	Protein	eQTL	Motif
		within 10 kb					
4	112769004	8410	rs385040	3a	CTCF		
					RAD21		
Chr	Coordinate	Distance (bp):	rsID	Score	Protein	eQTL	Motif
		within 100 kb					
4	112843090	65676	rs72899903	3a	FOS		
4	112845202	67788	rs76743823	2b			Foxl1
							FOXJ2
							FOXJ3
					CTCF		
					FOS		
4	112846378	68964	rs17589101	3a			REST
					CTCF		
4	112858746	81332	rs7670652	2b			Dobox4
							Evi-1
4	112865123	87709	rs72899924	2b			GCM
4	112873874	96460	rs115874059	3a			FOXO3A
					POLR2A		
4	112876056	98642	rs75690152	3a			FOXJ2
					CEBPB		
Chr	Coordinate	Distance (bp):	rsID	Score	Protein	eQTL	Motif
		within 150 kb					
4	112881725	104311	rs886567	3a			NGFI-C
							Egr-2
							EGR1
							Zif268
					GATA3		
					POLR2A		
					TFAP2A		
					GATA1		
					SMARCA4		
4	112883200	105786	rs757507	3a	EP300		
					FOS		
							Bsx

The coordinate of the lead SNP (rs17044095) is 112,777,414. Coordinates are 1-based. Chr stands for chromosome.
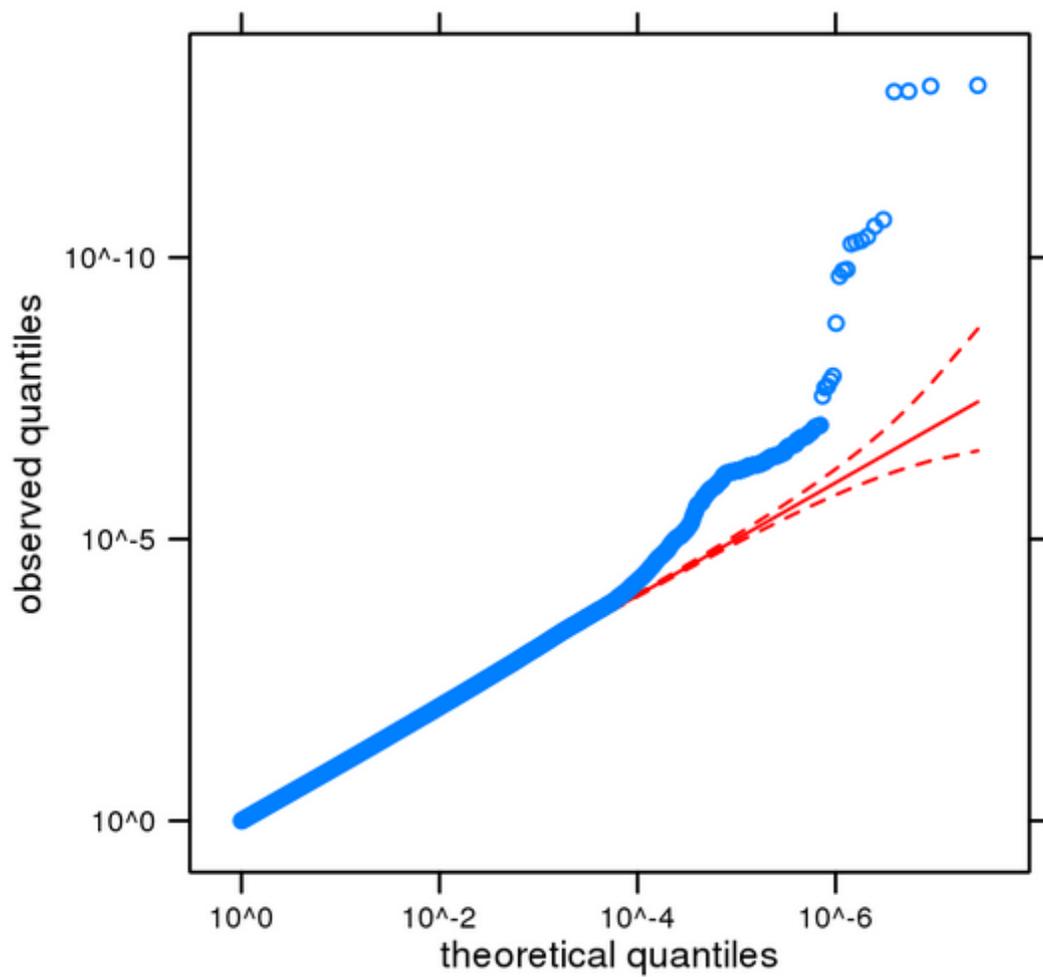
Figure 7. Quantile-Quantile Plot of *P*-values from Genome-Wide Association Study.

The solid red line symbolizes the expected *p*-values on a logarithmic scale under the null hypothesis. Blue dots represent observed values. The dashed red lines represent the 95% confidence envelope, assuming that the test results are independent.

Table 13. Quality Statistics for Index SNPs.

rsID	Average r^2	Minimum r^2	Allele Frequency
rs2129979	0.9859	0.9460	0.7039
rs192656182	0.7418	0.5297	0.0087
rs137882920	0.7190	0.6744	0.0171
rs2247036	0.9817	0.9682	0.4722
rs17044095	0.9964	0.9926	0.7699
rs117367662	0.7460	0.6936	0.0594
rs1650337	0.6620	0.5364	0.0002
rs75972139	0.9550	0.9371	0.9840
rs6445791	0.8707	0.8579	0.1666

Shown is information for the most-associated SNPs in each associated region for all 23andMe participants of European ancestry. All SNPs were imputed. Average r^2 is a measure of imputation quality, and minimum r^2 is a measure of consistency of imputation quality.
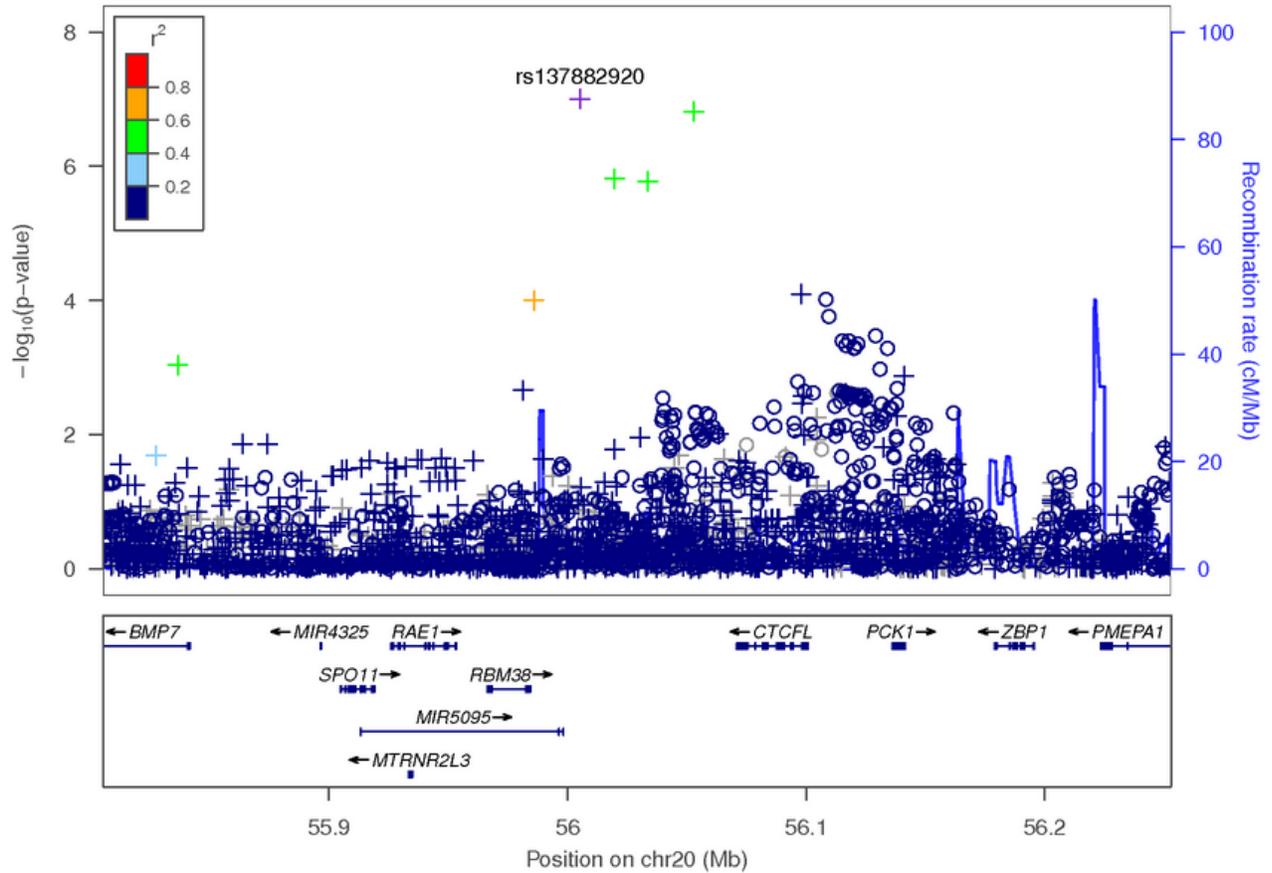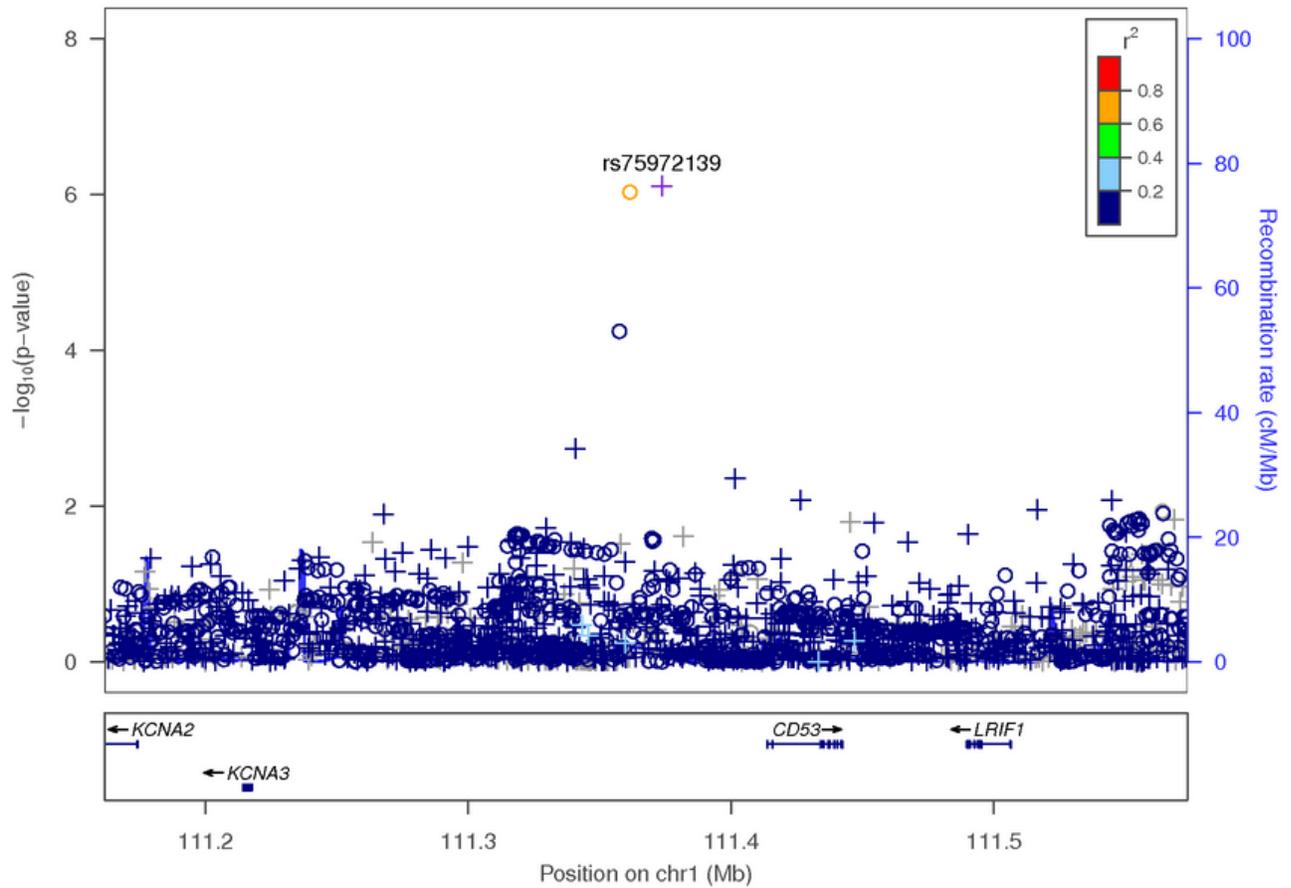
Figure 8. Regional Association Plot for rs137882920.

Association test results are shown as a distribution of position on chromosome 20 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. Genes in the vicinity of the lead SNP are shown along the X-axis.
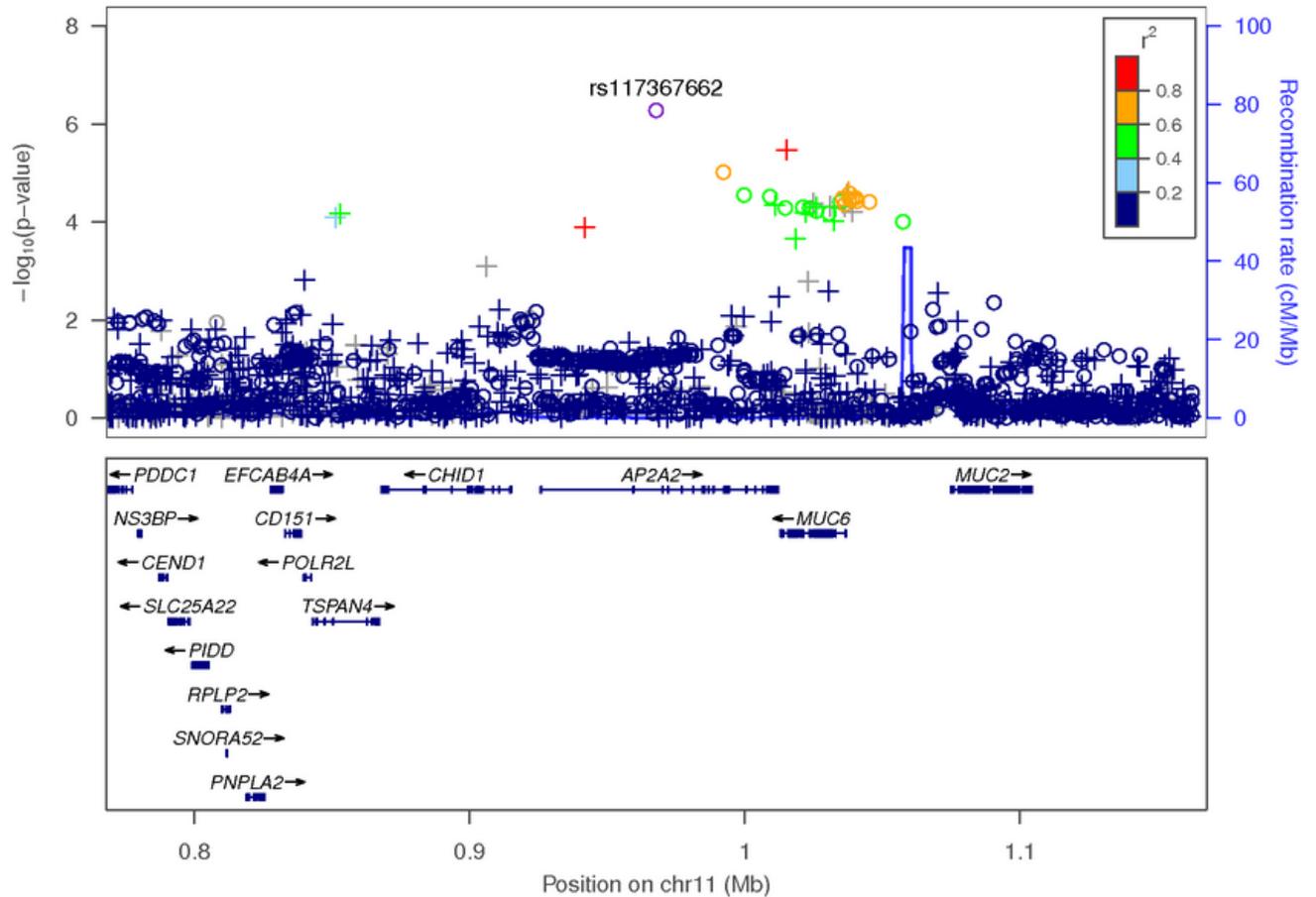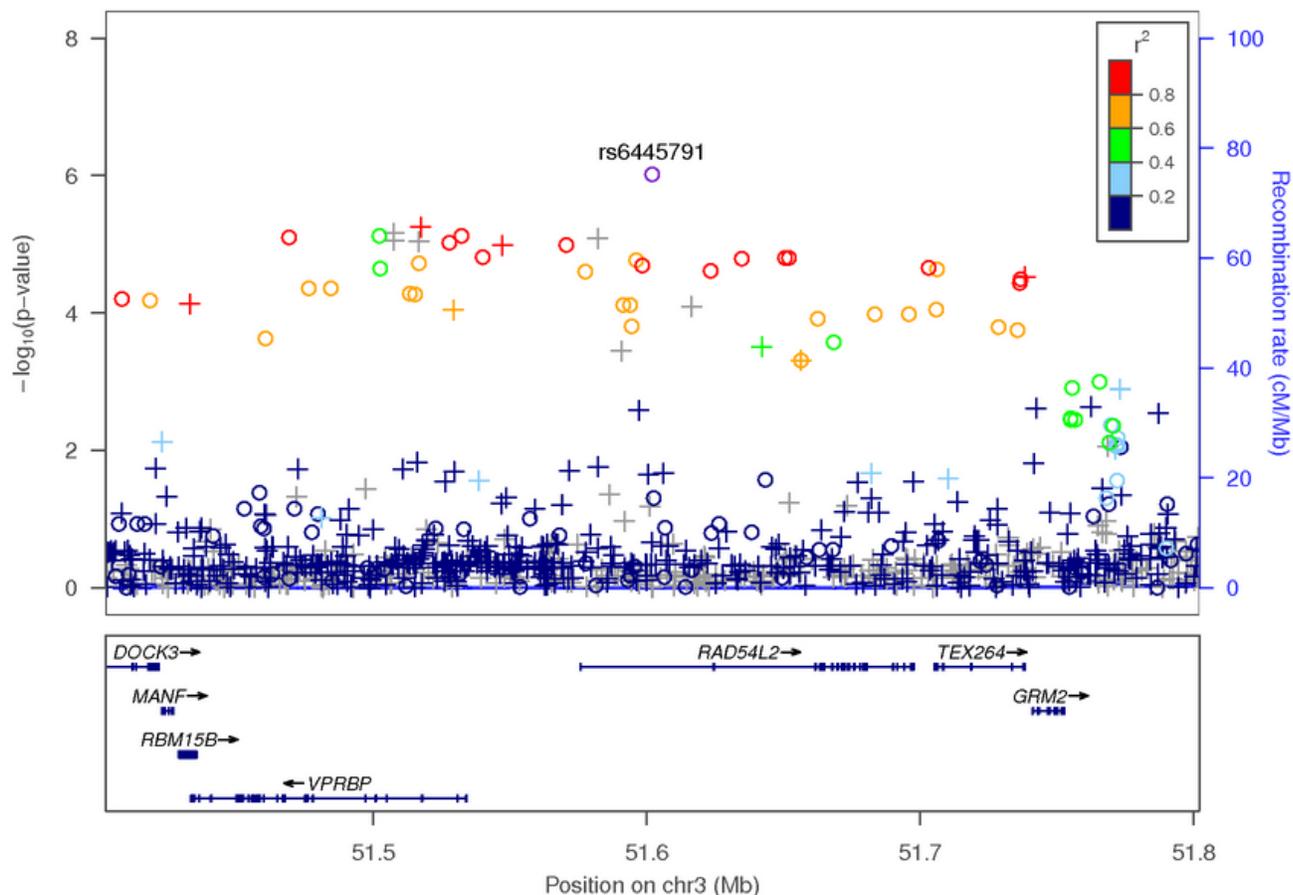
Figure 9. Regional Association Plot for rs117367662.

Association test results are shown as a distribution of position on chromosome 11 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. Genes in the vicinity of the lead SNP are shown along the X-axis.
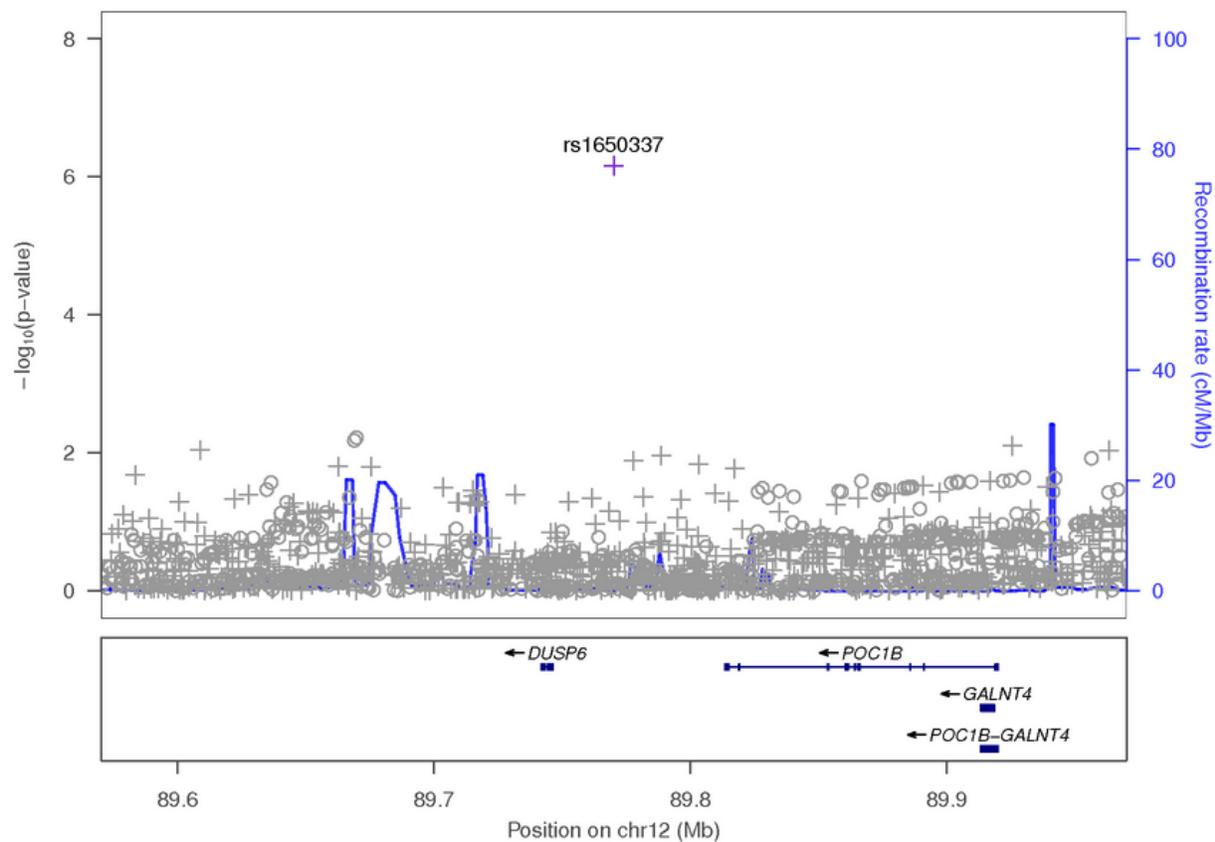
Figure 10. Regional Association Plot for rs1650337.

Association test results are shown as a distribution of position on chromosome 12 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. Genes in the vicinity of the lead SNP are shown along the X-axis.

Figure 11. Regional Association Plot for rs75972139.

Association test results are shown as a distribution of position on chromosome 1 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. Genes in the vicinity of the lead SNP are shown along the X-axis.

Figure 12. Regional Association Plot for rs6445791.

Association test results are shown as a distribution of position on chromosome 3 in the vicinity of the SNP along the X-axis versus negative transformed −log10 of association test P-values along the left Y-axis. The symbol "+" indicates a genotyped SNP; "o" indicates an imputed SNP. The colors indicate the strength of linkage disequilibrium (LD) with the index SNP. Blue lines indicate the recombination rate plotted along the right Y-axis, as a function of genomic position along the X-axis. Genes in the vicinity of the lead SNP are shown along the X-axis.

APPENDIX B: IRB APPROVAL LETTER

University of Pittsburgh
Institutional Review Board

Memorandum

To:	Dr. Michael Morowitz
From:	Sue Beers , Ph.D., Vice Chair
Date:	1/28/2014
IRB#:	PRO14010465
Subject:	Genetics of Appendicitis

The above-referenced protocol has been reviewed by the University of Pittsburgh Institutional Review Board. Based on the information provided to the IRB, this project includes no involvement of human subjects, according to the federal regulations [§45 CFR 46.102(f)]. That is, the investigator conducting research will not obtain information about research subjects via an interaction with them, nor will the investigator obtain identifiable private information. Should that situation change, the investigator must notify the IRB immediately.

Given this determination, you may now begin your project.

Please note the following information:

- If any modifications are made to this project, use the "**Send Comments to IRB Staff**" process from the project workspace to request a review to ensure it continues to meet the determination.
- Upon completion of your project, be sure to finalize the project by submitting a "**Study Completed**" report from the project workspace.

Please be advised that your research study may be audited periodically by the University of Pittsburgh Research Conduct and Compliance Office.

BIBLIOGRAPHY

1. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome research.* Sep 2012;22(9):1790-1797.
2. Rosenthal SL, Barmada MM, Wang X, Demirci FY, Kamboh MI. Connecting the dots: potential of data integration to identify regulatory SNPs in late-onset Alzheimer's disease GWAS findings. *PloS one.* 2014;9(4):e95152.
3. Murphy JB. Two Thousand Operations for Appendicitis with Deductions from his Personal Experience. *The American Journal of the Medical Sciences.* August, 1904 1904;128(2):187-210.
4. Humes DJ, Simpson J. Acute Appendicitis. *BMJ.* Sep 9 2006 2006;333(7567):530-534.
5. Buckius MT, McGrath B, Monk J, Grim R, Bell T, Ahuja V. Changing epidemiology of acute appendicitis in the United States: study period 1993-2008. *The Journal of surgical research.* Jun 15 2012;175(2):185-190.
6. Witt WP, Weiss, A.J., Elixhauser, A. Overview of Hospital Stays for Children in the United States, 2012. *Healthcare Cost and Utilization Project. Agency for Healthcare Research and Quality, Rockville, MD.* Dec 2014 2014;187.
7. Anderson JE, Bickler SW, Chang DC, Talamini MA. Examining a common disease with unknown etiology: trends in epidemiology and surgical management of appendicitis in California, 1995-2009. *World journal of surgery.* Dec 2012;36(12):2787-2794.
8. Cakmak YO, Ergelen R, Ekinci G, Kaspar EC. The short appendix vermiformis as a risk factor for colorectal cancer. *Clinical anatomy (New York, N.Y.).* Apr 2014;27(3):498-502.
9. Berry RJ. The True Caecal Apex, or the Vermiform Appendix: Its Minute and Comparative Anatomy. *Journal of anatomy and physiology.* Oct 1900;35(Pt 1):83-100.109.
10. Sonnenburg JL, Angenent LT, Gordon JI. Getting a grip on things: how do communities of bacterial symbionts become established in our intestine? *Nature immunology.* Jun 2004;5(6):569-573.
11. Bollinger RR, Everett ML, Palestrant D, Love SD, Lin SS, Parker W. Human secretory immunoglobulin A may contribute to biofilm formation in the gut. *Immunology.* Aug 2003;109(4):580-587.
12. Laurin M, Everett ML, Parker W. The cecal appendix: one more immune component with a function disturbed by post-industrial culture. *Anatomical record (Hoboken, N.J. : 2007).* Apr 2011;294(4):567-579.
13. Kaemmerer E, Schneider U, Klaus C, et al. Increased levels of deleted in malignant brain tumours 1 (DMBT1) in active bacteria-related appendicitis. *Histopathology.* Mar 2012;60(4):561-569.
14. Zhong D, Brower-Sinning R, Firek B, Morowitz MJ. Acute appendicitis in children is associated with an abundance of bacteria from the phylum Fusobacteria. *Journal of pediatric surgery.* Mar 2014;49(3):441-446.

15. Swidsinski A, Dorffel Y, Loening-Baucke V, et al. Mucosal invasion by fusobacteria is a common feature of acute appendicitis in Germany, Russia, and China. *Saudi journal of gastroenterology : official journal of the Saudi Gastroenterology Association.* Jan-Feb 2012;18(1):55-58.

16. Swidsinski A, Dorffel Y, Loening-Baucke V, et al. Acute appendicitis is characterised by local invasion with Fusobacterium nucleatum/necrophorum. *Gut.* Jan 2011;60(1):34-40.

17. Signat B, Roques C, Poulet P, Duffaut D. Fusobacterium nucleatum in periodontal health and disease. *Current issues in molecular biology.* 2011;13(2):25-36.

18. Pisacane A, de Luca U, Impagliazzo N, Russo M, De Caprio C, Caracciolo G. Breast feeding and acute appendicitis. *Bmj.* Apr 1 1995;310(6983):836-837.

19. St. Peter SD. Appendicitis. *Ashcraft's Pediatric Surgery, Fifth Ed.* 2010:549-556.

20. Carr NJ. The pathology of acute appendicitis. *Annals of diagnostic pathology.* Feb 2000;4(1):46-58.

21. Singh JP, Mariadason JG. Role of the faecolith in modern-day appendicitis. *Annals of the Royal College of Surgeons of England.* Jan 2013;95(1):48-51.

22. Chandrasegaram MD, Rothwell LA, An EI, Miller RJ. Pathologies of the appendix: a 10-year review of 4670 appendicectomy specimens. *ANZ journal of surgery.* Nov 2012;82(11):844-847.

23. Chang AR. An analysis of the pathology of 3003 appendices. *The Australian and New Zealand journal of surgery.* Apr 1981;51(2):169-178.

24. Arnbjornsson E, Bengmark S. Role of obstruction in the pathogenesis of acute appendicitis. *American journal of surgery.* Mar 1984;147(3):390-392.

25. Adamidis D, Roma-Giannikou E, Karamolegou K, Tselalidou E, Constantopoulos A. Fiber intake and childhood appendicitis. *International journal of food sciences and nutrition.* May 2000;51(3):153-157.

26. Barker DJ, Morris J. Acute appendicitis, bathrooms, and diet in Britain and Ireland. *British medical journal (Clinical research ed.).* Apr 2 1988;296(6627):953-955.

27. Barker DJ, Osmond C, Golding J, Wadsworth ME. Acute appendicitis and bathrooms in three samples of British children. *British medical journal (Clinical research ed.).* Apr 2 1988;296(6627):956-958.

28. Aravindan KP. Eosinophils in acute appendicitis: possible significance. *Indian journal of pathology & microbiology.* Oct 1997;40(4):491-498.

29. Santosh G, Aravindan KP. Evidence for eosinophil degranulation in acute appendicitis. *Indian journal of pathology & microbiology.* Apr-Jun 2008;51(2):172-174.

30. Woodruff SA, Masterson JC, Fillon S, Robinson ZD, Furuta GT. Role of eosinophils in inflammatory bowel and gastrointestinal diseases. *Journal of pediatric gastroenterology and nutrition.* Jun 2011;52(6):650-661.

31. Murphy CG, Glickman JN, Tomczak K, et al. Acute appendicitis is characterized by a uniform and highly selective pattern of inflammatory gene expression. *Mucosal immunology.* Jul 2008;1(4):297-308.

32. Andreu Ballester JC, Ballester F, Colomer Rubio E, Millan Scheiding M. Association between tonsillectomy, adenoidectomy, and appendicitis. *Revista espanola de enfermedades digestivas : organo oficial de la Sociedad Espanola de Patologia Digestiva.* Mar 2005;97(3):179-186.

33. Meljnikov I, Radojcic B, Grebeldinger S, Radojcic N. [History of surgical treatment of appendicitis]. *Medicinski pregled.* Sep-Oct 2009;62(9-10):489-492.

34.   Flum DR, McClure TD, Morris A, Koepsell T. Misdiagnosis of appendicitis and the use of diagnostic imaging. *Journal of the American College of Surgeons.* Dec 2005;201(6):933-939.

35.   Abou-Nukta F, Bakhos C, Arroyo K, et al. Effects of delaying appendectomy for acute appendicitis for 12 to 24 hours. *Archives of surgery (Chicago, Ill. : 1960).* May 2006;141(5):504-506; discussioin 506-507.

36.   Bickell NA, Aufses AH, Jr., Rojas M, Bodian C. How time affects the risk of rupture in appendicitis. *Journal of the American College of Surgeons.* Mar 2006;202(3):401-406.

37.   Basta M, Morton NE, Mulvihill JJ, Radovanovic Z, Radojicic C, Marinkovic D. Inheritance of acute appendicitis: familial aggregation and evidence of polygenic transmission. *American journal of human genetics.* Feb 1990;46(2):377-382.

38.   Perry T. KCE. Thirty-nine cases of appendicitis in a single family tree. *American journal of surgery.* 1936;46(2):259-265.

39.   Duffy DL, Martin NG, Mathews JD. Appendectomy in Australian twins. *American journal of human genetics.* Sep 1990;47(3):590-592.

40.   Sadr Azodi O, Andren-Sandberg A, Larsson H. Genetic and environmental influences on the risk of acute appendicitis in twins. *The British journal of surgery.* Nov 2009;96(11):1336-1340.

41.   Oldmeadow C, Mengersen K, Martin N, Duffy DL. Heritability and linkage analysis of appendicitis utilizing age at onset. *Twin research and human genetics : the official journal of the International Society for Twin Studies.* Apr 2009;12(2):150-157.

42.   Rivera-Chavez FA, Peters-Hybki DL, Barber RC, et al. Innate immunity genes influence the severity of acute appendicitis. *Annals of surgery.* Aug 2004;240(2):269-277.

43.   Daig R, Andus T, Aschenbrenner E, Falk W, Scholmerich J, Gross V. Increased interleukin 8 expression in the colon mucosa of patients with inflammatory bowel disease. *Gut.* Feb 1996;38(2):216-222.

44.   Mazzucchelli L, Hauser C, Zgraggen K, et al. Expression of interleukin-8 gene in inflammatory bowel disease is related to the histological grade of active inflammation. *The American journal of pathology.* May 1994;144(5):997-1007.

45.   Grimm MC, Elsbury SK, Pavli P, Doe WF. Interleukin 8: cells of origin in inflammatory bowel disease. *Gut.* Jan 1996;38(1):90-98.

46.   Dalal I, Somekh E, Bilker-Reich A, Boaz M, Gorenstein A, Serour F. Serum and peritoneal inflammatory mediators in children with suspected acute appendicitis. *Archives of surgery (Chicago, Ill. : 1960).* Feb 2005;140(2):169-173.

47.   Zeillemaker AM, Hoynck van Papendrecht AA, Hart MH, Roos D, Verbrugh HA, Leguit P. Peritoneal interleukin-8 in acute appendicitis. *The Journal of surgical research.* May 1996;62(2):273-277.

48.   Paajanen H, Mansikka A, Laato M, Ristamaki R, Pulkki K, Kostiainen S. Novel serum inflammatory markers in acute appendicitis. *Scandinavian journal of clinical and laboratory investigation.* 2002;62(8):579-584.

49.   Yoon DY, Chu J, Chandler C, Hiyama S, Thompson JE, Hines OJ. Human cytokine levels in nonperforated versus perforated appendicitis: molecular serum markers for extent of disease? *The American surgeon.* Dec 2002;68(12):1033-1037.

50.   Barker DJ, Morris JA, Simmonds SJ, Oliver RH. Appendicitis epidemic following introduction of piped water to Anglesey. *Journal of epidemiology and community health.* Jun 1988;42(2):144-148.

51. Iyengar SK, Elston RC. The genetic basis of complex traits: rare variants or "common gene, common disease"? *Methods in molecular biology (Clifton, N.J.).* 2007;376:71-84.

52. Mason RJ. Surgery for appendicitis: is it necessary? *Surgical infections.* Aug 2008;9(4):481-488.

53. Barrett M.L. HAL, Andrews R.M. Trends in Rates of Perforated Appendix, 2001–2010. Statistical Brief #159. *Healthcare Cost and Utilization Project. Agency for Healthcare Research and Quality, Rockville, MD.* July 2013 2013.

54. Tsay JH, Lee CH, Hsu YJ, et al. Disparities in appendicitis rupture rate among mentally ill patients. *BMC public health.* 2007;7:331.

55. Hsia R.Y. KAH, Srebotnjak T., Maselli J. Health Care as a "Market Good"? Appendicitis as a Case Study. *Arch Intern Med.* May 28 2012 2012;172(10):818-819.

56. Kastenberg ZJ, Hurley MP, Luan A, et al. Cost-effectiveness of preoperative imaging for appendicitis after indeterminate ultrasonography in the second or third trimester of pregnancy. *Obstetrics and gynecology.* Oct 2013;122(4):821-829.

57. Guttman R, Goldman RD, Koren G. Appendicitis during pregnancy. *Canadian family physician Medecin de famille canadien.* Mar 2004;50:355-357.

58. Bolin TD, Wong S, Crouch R, Engelman JL, Riordan SM. Appendicectomy as a therapy for ulcerative proctitis. *The American journal of gastroenterology.* Oct 2009;104(10):2476-2482.

59. Koutroubakis IE, Vlachonikolis IG. Appendectomy and the development of ulcerative colitis: results of a metaanalysis of published case-control studies. *The American journal of gastroenterology.* Jan 2000;95(1):171-176.

60. Wagner PL, Eachempati SR, Soe K, Pieracci FM, Shou J, Barie PS. Defining the current negative appendectomy rate: for whom is preoperative computed tomography making an impact? *Surgery.* Aug 2008;144(2):276-282.

61. Seetahal SA, Bolorunduro OB, Sookdeo TC, et al. Negative appendectomy: a 10-year review of a nationally representative sample. *American journal of surgery.* Apr 2011;201(4):433-437.

62. Chang AL, Raber I, Xu J, et al. Assessment of the genetic basis of rosacea by genome-wide association study. *The Journal of investigative dermatology.* Jun 2015;135(6):1548-1555.

63. Ferreira MA, Matheson MC, Tang CS, et al. Genome-wide association analysis identifies 11 risk variants associated with the asthma with hay fever phenotype. *The Journal of allergy and clinical immunology.* Dec 30 2013.

64. Apfel CC, Heidrich FM, Jukar-Rao S, et al. Evidence-based analysis of risk factors for postoperative nausea and vomiting. *British journal of anaesthesia.* Nov 2012;109(5):742-753.

65. Eriksson N, Macpherson JM, Tung JY, et al. Web-based, participant-driven studies yield novel genetic associations for common traits. *PLoS genetics.* Jun 2010;6(6):e1000993.

66. Abecasis GR, Altshuler D, Auton A, et al. A map of human genome variation from population-scale sequencing. *Nature.* Oct 28 2010;467(7319):1061-1073.

67. Browning SR, Browning BL. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *American journal of human genetics.* Nov 2007;81(5):1084-1097.

68.   Howie B, Fuchsberger C, Stephens M, Marchini J, Abecasis GR. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics.* Aug 2012;44(8):955-959.

69.   Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics.* Aug 2003;164(4):1567-1587.

70.   Henn BM, Hon L, Macpherson JM, et al. Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PloS one.* 2012;7(4):e34267.

71.   An integrated encyclopedia of DNA elements in the human genome. *Nature.* Sep 6 2012;489(7414):57-74.

72.   Boyle AP, Araya CL, Brdlik C, et al. Comparative analysis of regulatory information and circuits across distant species. *Nature.* Aug 28 2014;512(7515):453-456.

73.   Xie D, Boyle AP, Wu L, Zhai J, Kawli T, Snyder M. Dynamic trans-acting factor colocalization in human cells. *Cell.* Oct 24 2013;155(3):713-724.

74.   Sulu B. Demographic and Epidemiologic Features of Acute Appendicitis. *Appendicitis - A Collection of Essays From Around the World.* InTech2012.

75.   Ashley DJ. Observations on the epidemiology of appendicitis. *Gut.* Dec 1967;8(6):533-538.

76.   Addiss DG, Shaffer N, Fowler BS, Tauxe RV. The epidemiology of appendicitis and appendectomy in the United States. *American journal of epidemiology.* Nov 1990;132(5):910-925.

77.   Degner JF, Pai AA, Pique-Regi R, et al. DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature.* Feb 16 2012;482(7385):390-394.

78.   Poli V. The role of C/EBP isoforms in the control of inflammatory and native immunity functions. *The Journal of biological chemistry.* Nov 6 1998;273(45):29279-29282.

79.   Ruther U, Wagner EF, Muller R. Analysis of the differentiation-promoting potential of inducible c-fos genes introduced into embryonal carcinoma cells. *The EMBO journal.* Jul 1985;4(7):1775-1781.

80.   Wagner EF. Bone development and inflammatory disease is regulated by AP-1 (Fos/Jun). *Annals of the rheumatic diseases.* Jan 2010;69 Suppl 1:i86-88.

81.   Bossis G, Malnou CE, Farras R, et al. Down-regulation of c-Fos/c-Jun AP-1 dimer activity by sumoylation. *Molecular and cellular biology.* Aug 2005;25(16):6964-6979.

82.   Franceschini A, Szklarczyk D, Frankild S, et al. STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research.* Jan 2013;41(Database issue):D808-815.

83.   Lee I, Blom UM, Wang PI, Shim JE, Marcotte EM. Prioritizing candidate disease genes by network-based boosting of genome-wide association data. *Genome research.* Jul 2011;21(7):1109-1121.

84.   Matsusaka T, Fujikawa K, Nishio Y, et al. Transcription factors NF-IL6 and NF-kappa B synergistically activate transcription of the inflammatory cytokines, interleukin 6 and interleukin 8. *Proceedings of the National Academy of Sciences of the United States of America.* Nov 1 1993;90(21):10193-10197.

85.   Stein B, Baldwin AS, Jr. Distinct mechanisms for regulation of the interleukin-8 gene involve synergism and cooperativity between C/EBP and NF-kappa B. *Molecular and cellular biology.* Nov 1993;13(11):7191-7198.

86. Hoffmann E, Dittrich-Breiholz O, Holtmann H, Kracht M. Multiple control of interleukin-8 gene expression. *Journal of leukocyte biology*. Nov 2002;72(5):847-855.

87. Mesbah-Uddin M, Elango R, Banaganapalli B, Shaik NA, Al-Abbasi FA. In-silico analysis of inflammatory bowel disease (IBD) GWAS loci to novel connections. *PloS one*. 2015;10(3):e0119420.

88. Fact Sheet: President Obama's Precision Medicine Initiative [press release]. The White HouseJanuary 30, 2015.

89. Bielinski SJ, Olson JE, Pathak J, et al. Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time-using genomic data to individualize treatment protocol. *Mayo Clinic proceedings*. Jan 2014;89(1):25-33.

90. Health Information Exchange (HIE). *Office of the National Coordinator for Health Information Technology*. May 12, 2014.

91. Nielsen DE, El-Sohemy A. Disclosure of genetic information and change in dietary intake: a randomized controlled trial. *PloS one*. 2014;9(11):e112665.

92. Hyman RW, Fukushima M, Jiang H, et al. Diversity of the vaginal microbiome correlates with preterm birth. *Reproductive sciences (Thousand Oaks, Calif.)*. Jan 2014;21(1):32-40.

93. Vu BG, Stach CS, Salgado-Pabon W, Diekema DJ, Gardner SE, Schlievert PM. Superantigens of Staphylococcus aureus from patients with diabetic foot ulcers. *The Journal of infectious diseases*. Dec 15 2014;210(12):1920-1927.

94. Devaraj S, Hemarajata P, Versalovic J. The human gut microbiome and body metabolism: implications for obesity and diabetes. *Clinical chemistry*. Apr 2013;59(4):617-628.

95. Sato J, Kanazawa A, Ikeda F, et al. Gut dysbiosis and detection of "live gut bacteria" in blood of Japanese patients with type 2 diabetes. *Diabetes care*. Aug 2014;37(8):2343-2350.

96. Naccache SN, Federman S, Veeraraghavan N, et al. A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples. *Genome research*. Jul 2014;24(7):1180-1192.

97. Strauss J, Kaplan GG, Beck PL, et al. Invasive potential of gut mucosa-derived Fusobacterium nucleatum positively correlates with IBD status of the host. *Inflammatory bowel diseases*. Sep 2011;17(9):1971-1978.

98. Wang N, Guo B, Liu Y. Methods and materials for quantification of fusobacterium nucleatum dna in stool to diagnose colorectal neoplasm. Google Patents; 2014.

99. Loria K. This Woman's Revolutionary Idea Made Her A Billionaire — And Could Change Medicine. *Business Insider*. Sept 29, 2014.

100. Diamandis EP. Theranos phenomenon: promises and fallacies. *Clinical chemistry and laboratory medicine : CCLM / FESCC*. Jun 1 2015;53(7):989-993.

101. Potgieter M, Bester J, Kell DB, Pretorius E. The dormant blood microbiome in chronic, inflammatory diseases. *FEMS microbiology reviews*. Jul 2015;39(4):567-591.

102. Colman RJ, Rubin DT. Fecal microbiota transplantation as therapy for inflammatory bowel disease: a systematic review and meta-analysis. *Journal of Crohn's & colitis*. Dec 2014;8(12):1569-1581.

103. Mayo Clinic and Whole Biome Announce Collaboration: Joint Development of Microbiome Diagnostic Testing to Focus on Women's Health and Preterm Labor [press release]. Mayo Clinic News Network: Mayo Clinic, May 20, 2014.