2015

Identification of Two Novel Genome-Wide Significant Single Nucleotide Polymorphisms, associated with Barrett's Oesophagus, determined by further Replication of a Genome-Wide Association Study

Chegwidden, Laura

1. COPYRIGHT STATEMENT

Identification of Two Novel Genome-Wide Significant Single Nucleotide Polymorphisms, associated with Barrett's Oesophagus, determined by further Replication of a Genome-Wide Association Study

by

LAURA CHEGWIDDEN

A thesis submitted to Plymouth University

for the degree of

MASTER OF PHILOSOPHY

Plymouth University Peninsula Schools of Medicine and Dentistry

November 2014

Laura Chegwidden

Identification of Two Novel Genome-Wide Significant Single Nucleotide Polymorphisms, associated with Barrett's Oesophagus, determined by further Replication of a Genome-Wide Association Study

2. ABSTRACT

Barrett's oesophagus (BE) is a common premalignant condition to oesophageal adenocarcinoma (EAC). A previous genome-wide association study (GWAS) identified BE susceptibility Single Nucleotide Polymorphisms (SNPs) on chromosome 6p21, within the HLA region, and16q23, where the closest protein-coding gene was *FOXF1*.

The replication study outlined in this thesis aimed to identify possible additional variants that did not reach genome-wide significance in the GWAS, in up to 10,158 BE patients and 21,062 controls. Meta-analysis of the data identified two further BE susceptibility SNPs: rs3072 (2p24.1; OR=1.14; 95%CI 1.09-1.18; $P=1.8\times10^{-11}$); and rs2701108 (12q24.21; OR=0.90; 95%CI 0.86-0.93; $P=7.5\times10^{-9}$). The two closest protein-coding genes, and most likely functional targets, are the bone morphogenetic protein pathway ligand *GDF7* (rs3072) and *TBX5* (rs2701108).

A second GWAS of combined BE and EAC cases was recently published, analysing a total of 922,031 SNPs, where 87 of 94 associated SNPs with $P<1\times10^{-4}$ were selected for further replication, identified four SNPs (three loci) with BE/EAC risk in *CRTC1* and *BARX1* and within 100kb of *FOXP1*. Our data supported three of the BE/EAC-associated SNPs and meta-analysis of all 87 SNPs detected a further susceptibility locus, rs3784262, near *ALDH1A2* (OR=0.90, 95%CI 0.87-0.93, $P=3.72\times10^{-9}$).

Overall, two novel BE susceptibility loci have been identified and data has been provided to support three previously identified BE/EAC SNPs and one additional BE/EAC locus. To date, genes implicated in BE susceptibility appear to encode transcription factors involved in thoracic, diaphragmatic and oesophageal development or inflammatory response proteins.

3. CONTENTS

4. LIST OF TABLES AND FIGURES

4.1 LIST OF FIGURES

4.2 LIST OF TABLES

5. ACKNOWLEDGEMENTS

I would like to express my gratitude to Professor Janusz Jankowski for giving me the opportunity to work on the ChOPIN and AspECT clinical trials at Queen Mary University of London and the University of Plymouth. I would also like to thank Janusz for permitting the use of his samples in this study and his continued support and advice throughout. I would also like to thank Dr Elaine Green and Professor Simon Jackson for their invaluable constructive criticism and advice during the project, particularly in the final stages.

My thanks go to Professor Ian Tomlinson for allowing me to work in his lab at the Wellcome Trust Centre for Human Genetics, Oxford. A special thank you goes to Dr Claire Palles, from whom I learnt a lot within just five weeks! Thank you Claire, for your support and guidance, particularly whilst genotyping UKREP3 and the Levine et al (2013) [3] SNPs and also for introducing me to the related statistical analyses. Thank you to my colleagues at Plymouth University; Dr Claire Adams and Dr Garry Farnham for their continued support and advice throughout my time at the University of Plymouth. Also, thank you to Dr Xinzhong Li for advice regarding statistical analyses. Data for this study replied upon nationwide and worldwide collaborations between healthcare professionals, scientists and patients. I would therefore like to thank all authors on the BE GWAS paper (Su et al (2012)) and the recent replication paper (Palles et al (2015)). Thank you to all ChOPIN participants and clinical trial teams (Principle Investigators and nurses) for their continued support of BE research. A special thank you to Kirstie Walker, research nurse at North Tyneside General Hospital; I will miss the phone calls and emails we shared throughout my time working on the clinical trials!

Finally, but by no means least, I would like to thank my parents. Without their unconditional love and support I would not be where I am today, a massive thank you to you both!

6. AUTHOR'S DECLARATION AND WORD COUNT

At no time during the registration for the degree of Master of Philosophy has the author been registered for any other University award without prior agreement of the Graduate Committee.

Work submitted for this research degree at Plymouth University Peninsula Schools of Medicine and Dentistry has not formed part of any other degree either at Plymouth University Peninsula Schools of Medicine and Dentistry or at another establishment.

This study was financed with the aid of a studentship from Plymouth University Peninsula Schools of Medicine and Dentistry and carried out in collaboration with Wellcome Trust Centre for Human Genetics, Oxford.

A programme of advanced study was undertaken, which included Good Clinical Practice training and two six-week online bioinformatics courses.

Publications:

Palles, C., et al., Polymorphisms near TBX5 and GDF7 are associated with increased risk for Barrett's esophagus. Gastroenterology, 2015. 148(2): p. 367-78.

Presentation and Conferences Attended:

PU PSMD Research Event (Poster Presentation) - Wednesday 22nd October 2014

Word count of main body of thesis: 20,842

Sign ……..

Date ……..

7. INTRODUCTION

Barrett's Oesophagus (BE; OMIM 614266) is a common premalignant condition, affecting up to 2% of the general population in the Western world [7], and arises from the transition of cells from normal squamous epithelium to columnar epithelium. It was first described by Dr Norman Rupert Barrett in 1950 [8]. BE is a precursor to Oesophageal Adenocarcinoma (EAC) and has been shown to follow a metaplasia (BE) – dysplasia (presence of abnormal cells within BE) - adenocarcinoma sequence (MCS) [9-11]. One risk factor for BE is reflux, potentially leading to oesophagitis (inflammation of the oesophagus). It is predicted that 10% of those with oesophagitis will go on to develop BE, where short segment BE (SSBE; <3cm) is more prevalent (8-20%) than long segment BE (LSBE; >3cm), with just 1% prevalence. BE patients have a 2-24% risk of developing high grade dysplasia (HGD), and a 2-5% risk of developing EAC [11]. Whilst patients with BE and HGD have a 40-50% risk of developing EAC within 5 years [12]. The incidence of EAC has been rising by 3% each year for the last 30 years and is the fifth most common cancer in the UK [13].

7.1 THE OESOPHAGUS

The oesophagus is a muscular tube, measuring between 18-25cm in humans, through which food passes from the mouth to the stomach, aided by peristalsis. Two sphincters are present in the upper and lower ends of the oesophagus (upper oesophageal sphincter (UOS) and lower oesophageal sphincter (LOS)). The two sphincters act as barriers and prevent backflow of food and/or stomach acid. Food passes through the oesophagus, past the gastro-oesophageal junction (GEJ) and into the stomach. The wall of the oesophagus consists of four main layers: mucosa, submucosa, muscularis propria, and adventitia (Figure 7.1). The mucosa is composed of the stratified squamous epithelium, lamina propria and muscularis mucosa. The submucosa contains bloods vessels, nerves and glands. The muscularis propria consists of skeletal (upper muscularis propria) and smooth (lower muscularis propria) muscle

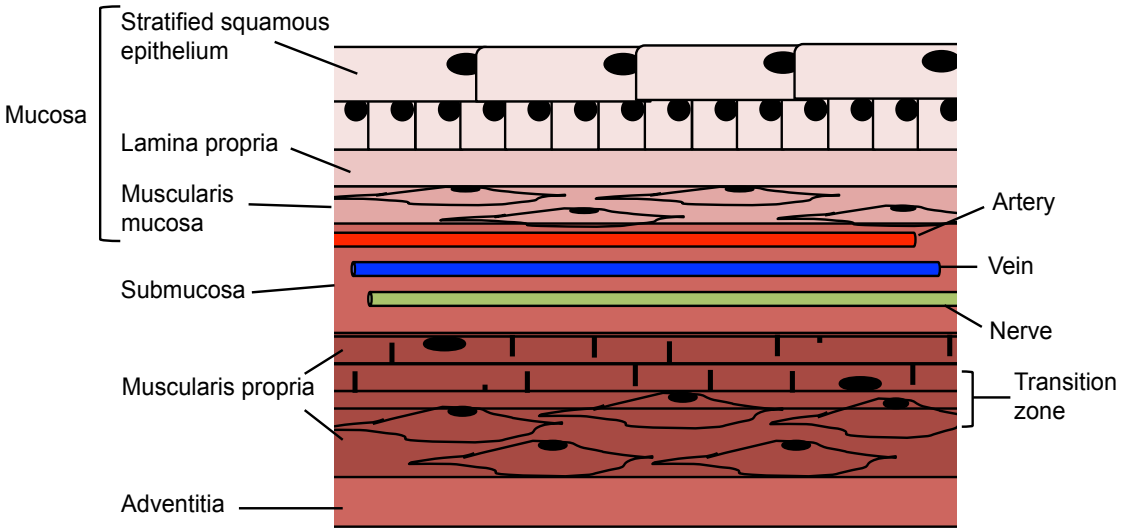


Figure 7.1: Oesophageal Wall Structure. The oesophageal wall is composed of four main sections: mucosa, submucosa, muscularis propria and adventitia. The mucosa contains the stratified squamous epithelium, lamina propria and muscularis mucosa. The submucosa contains bloods vessels, nerves and glands. The muscularis propria consists of skeletal (upper) and smooth (lower) muscle fibres; there is combination of both muscle types in the transition zone. The adventitia is an external fibrous layer composed of loose connective tissue.

fibres; where grouping of both muscle types is known as the transition zone. The adventitia is an external fibrous layer composed of loose connective tissue. Oesophageal mucosa is pale pink in colour in contrast with the darker gastric mucosa present in the stomach. The Z-line is the term used to describe the position at which the two mucosa meet. The oesophagus does not have a serous membrane, unlike the rest of the gastrointestinal (GI) tract. It is thought that the absence of this layer allows oesophageal cancers to spread more easily.

7.2 PREVALENCE OF BE

There is a large variation in the prevalence of BE across the globe, particularly when comparing Western to Eastern worlds. The majority of data collected for BE prevalence is based on patients undergoing endoscopy procedures, and so are not a true reflection of prevalence, as BE cases can be asymptomatic [14]. However, some studies have based their research on a random selection of the general population; Hayeck et al (2010) estimated the prevalence of BE in the USA at 5.6% [15], Ronkainen et al (2005) estimated BE prevalence in Sweden at 1.6% [16], and Zagari et al (2008) estimated the prevalence of BE in Italy at 1.3% [17]. The true prevalence of BE worldwide is unknown due to asymptomatic cases and a lack of resources in some

countries, however it has been estimated between 1.6-3% [10].

7.3 BE DIAGNOSIS

BE is initially diagnosed by endoscopy, which allows the measurement of the length of BE segment. A change in colour of the lining of the lower oesophagus from its normal pale pink (stratified squamous epithelium) to a red colour (columnar epithelium) suggests that BE has developed. The junction of these two epithelia is called the Z-line. If this colour change is present then biopsies will be taken from various sites in the oesophagus. The biopsies will be histologically examined for the characteristic columnar cells to confirm diagnosis. When examining the biopsies, the histologist will also look for the presence of goblet cells (known as Intestinal Metaplasia; IM) and any signs of dysplasia (the presence of abnormal, precancerous cells).

The degree of dysplasia in BE can be classed as one of the following:

1. Negative for dysplasia; presence of mild abnormalities (e.g. nuclear enlargement, crowding) usually confined to the lower glands, which can be misinterpreted as dysplasia.
2. Indefinite dysplasia; where the significance of any observed dysplasia is uncertain. For example, the nuclei may be enlarged but uniform in size and shape, unlike that seen in dysplasia.
3. Low grade dysplasia (LGD); the presence of dysplastic nuclei that are located in the base of the cell.
4. High grade dysplasia (HGD); presence of inconsistent dysplastic nuclei with loss of nuclear polarity confined to the mucosa, without crossing the basement membrane.
5. Intramucosal adenocarcinoma; where the dysplasia crosses the basement membrane and invades the lamina propria (see Figure 7.1).
6. Invasive adenocarcinoma; where dysplasia reaches deeper into the tissue and invades the muscularis mucosa (see Figure 7.1) [7, 18-20].

7.4 INCREASED RISK FACTORS ASSOCIATED WITH BE

Risk factors associated with BE include gender, age, ethnicity, smoking, obesity (waist-hip ratio; WHR) and gastro-oesophageal reflux disease (GERD).

7.4.1 Gender

BE is more common in males (with 2:1 male:female ratio), possibly due to a protective effect, related to sex-specific hormone seen in premenopausal women (possibly oestrogen) [18, 21, 22].

7.4.2 Age and Ethnicity

BE is more common in individuals over 50 years of age and in those of white Caucasian origin [23].

7.4.3 Smoking

There is conflicting evidence about the association between smoking and BE. Kubo et al (2009) [24] conclude that there was no association, however their sample size was small (320 BE cases, 316 GERD cases and 317 population controls) which limits the power to detect an association. More recently, Balasubramanian et al (2013) [25] tested the association between smoking and BE with a case sample size of 1,056 and concluded that smoking was an independent risk factor for BE, which significantly increases with the intensity of smoking and that stopping for 20 years or more reduces this risk.

7.4.4 Obesity – Body Mass Index (BMI) and Waist-Hip Ratio (WHR)

Similarly to smoking, there is some conflicting evidence over the association between obesity and BE, although the majority of studies do confirm a link between the two [26]. Kamat et al (2009) [27] showed that there was a small statistically significant association between obesity and BE. However, Kubo et al (2013) [28] concluded that there was not an association between increased BMI and BE, rather that there was an association with larger waist circumference. Similar studies have confirmed that increased waist-hip ratio (WHR) is a risk factor, rather than increased BMI, for

developing BE, particularly LSBE [29, 30]. It has been shown that as women reach menopause, they tend to gain weight, and so are at more risk of developing BE later in life compared to men [21].

7.4.5 Gastro-oesophageal Reflux Disease (GERD)

One of the main risk factors is GERD, which is usually caused by the failure of the lower oesophageal sphincter (LOS), resulting in stomach acid entering the oesophagus causing mucosal damage [31, 32].

Risk factors for GERD include hiatus hernia, where the upper part of the stomach protrudes into the thorax through a weakened diaphragm [33]; obesity; factors that increase gastrin production and acidity, for example Zollinger-Ellison syndrome [34] and hypercalcaemia [35]; and factors that result in oesophageal dysmotility, such as systemic sclerosis [36].

Diagnosis of GERD is usually given when the typical symptoms are present, such as heartburn and regurgitation [37]. A 24-hour pH-monitoring device may be used to confirm diagnosis, which sits inside the patient's oesophagus and records the pH level over a 24-hour period. An endoscopy may be requested if further investigation is needed.

It is important to control the amount of acid released into the oesophagus to prevent possible complications, such as oesophagitis (inflammation of the oesophageal lining) and BE. Treatment for GERD includes lifestyle changes, medication and surgery. Weight loss, quitting smoking and removal of alcohol from the diet appears to reduce reflux symptoms, as does moderate exercise and avoiding specific foods that increase stomach acid (such as coffee, chocolate and spicy foods). Medications include Proton Pump Inhibitors (PPIs), H2 receptor blockers and antacids [38, 39]. Nissen fundoplication, a surgical procedure, is an option for those who respond well to medication but don't want to continue it long-term or for those who have side effects to

medication (mainly diarrhea, fractures or drug interactions), and involves wrapping of the upper section of the stomach around the LOS to strengthen it.

Between 5-10% of patients with GERD go on to develop BE [6], indicating other genetic and/or environmental factors associated with BE [7, 24, 29, 40-47].

7.5 DECREASED RISK FACTORS ASSOCIATED WITH BE

Two basic factors associated with decreased risk of BE are consumption of a good diet and nutrients. Other factors associated with decreased BE risk are *Helicobacter pylori* (*H.pylori*) infection and the use of Aspirin and Non-Steroidal Anti-Inflammatory Drugs (NSAIDs). A large trial is currently assessing the long-term value of aspirin preventing BE developing into cancer.

7.5.1 Diet and Nutrient Intake

A study published by Kubo et al (2009) [42] found that higher intakes of omega-3-fatty-acids, polyunsaturated fat and fibre from various sources, including fruits and vegetables were all associated with a lower risk of BE. They also found that higher meat intakes were associated with a lower risk of long-segment BE. However, higher trans-fat (a type of unsaturated fat) intakes were associated with increased BE risk.

Nutrient intake has also been associated with a reduction in BE risk. Kubo et al (2008) [41] showed that high intakes of vitamin C, beta-carotene, and vitamin E were all inversely associated with BE risk, with the latter being the most significant association.

7.5.2 Helicobacter pylori Infection

H.pylori is a gram-negative microaerophilic bacterium found in the stomach, and is associated with decreased BE risk. It is estimated that >50% of the world's population carry *H.pylori* in their upper GI tract. The infection is more prevalent in developing countries, and is decreasing in Western countries. It is thought that the bacterium reduces BE risk via factors which appear to reduce gastric acid [48]. Various studies investigating the effects of *H.pylori* have shown that there is considerable decreased risk of BE in patients who carry *H.pylori*. A study completed by Corley et al (2008) [49]

showed that BE patients were less likely to have antibodies for *H.pylori* (OR=0.42) than population controls. They also showed that the inverse association was stronger among those with BMIs<25 (OR=0.03) compared to those with BMIs>30 (OR=0.43). Similarly, a study completed by Anderson et al (2008) [50] found that those positive for *H.pylori* were inversely associated with BE (OR=0.41). However, they also showed that the inverse associations between the presence of the bacterium and BE remained in patients who did not experience reflux symptoms, suggesting that the bacterium acts via more than one mechanism to reduce BE risk.

7.5.3 Aspirin and Non-Steroidal Anti-Inflammatory Drugs

There are conflicting reports about the protective effects of Aspirin and NSAID use. Anderson et al (2006) [51] showed that the use of these drugs were associated with a reduced risk of BE, with OR=0.53 for Aspirin and OR=0.40 for NSAIDs. However, a similar study published by Wang et al (2010) [52] suggested that the drugs might act after the formation of BE in the inflammation-metaplasia-dysplasia-adenocarcinoma sequence, and hence protect against EAC instead.

7.6 TREATMENT AND PROGNOSIS

7.6.1 Treatment

Patients with reflux symptoms will more than likely opt to take either over-the-counter antacids or prescribed PPIs. These drugs act by preventing cells in the lining of the stomach from producing too much acid. Treatment for BE depends on the stage of disease. Patients with non-dysplastic or LGD BE are more likely to undergo endoscopic surveillance, where an endoscopy procedure is carried out every two years with the aim to catch possible progression of the disease early [53]. There are surgical procedures available for BE patients who have LGD or HGD, including Nissen fundoplication, endoscopic mucosal resection (EMR), endoscopic ablative therapy (including photodynamic therapy (PDT), argon plasma coagulation (APC), cryotherapy, and radio-frequency ablation (RFA)), and oesophagectomy to name a few. Nissen fundoplication is a surgical procedure used to treat GERD and hiatus hernia and

involves wrapping the upper part of the stomach (fundus) around the lower oesophagus. The idea behind the procedure is to mimic the function of the LOS to prevent acid reflux. EMR involves the endoscopic removal of the affected tissue, which can then be histologically examined [54]. Endoscopic ablative therapy involves the destruction of the effected tissue, which is then usually followed by repopulation with normal squamous epithelia in a non-acidic environment [7, 19, 55, 56]. The issue with this technique is ensuring that all BE tissue is ablated. An oesophagectomy is usually employed if patients have HGD or EAC, and involves the removal of all or part of the oesophagus. However, it is associated with high mortality and morbidity rates [7, 19, 57].

7.6.2 Prognosis

It is important to limit the amount of acid entering the oesophagus in order to prevent an increase in BE segment size. This can be achieved via the treatment methods outlined above (section 7.6.1). An increase in BE segment length increases the chance of developing EAC. Only 2-5% of BE patients progress to EAC (~5% lifetime risk in men and ~3% in women), which has a high mortality rate [6, 7, 20, 58, 59]. One study has shown that BE patients die more commonly of bronchopneumonia and ischaemic heart disease rather than EAC [59]. BE cannot be cured unless all affected tissue has been removed. It has also been shown that BE can regress, although short segment BE is more likely to regress than long segment BE, as were those treated surgically [60, 61].

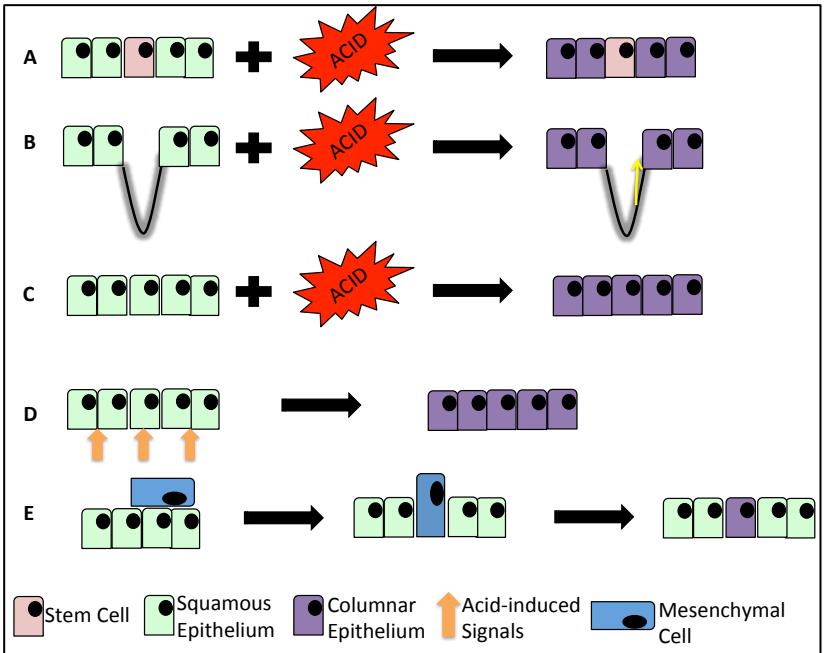


Figure 7.2: Hypotheses for BE Cellular Origin. (A) reprogramming of a stem cell, (B) Proliferation and migration of cells from the glands after acid exposure, (C) trans-differentiation of squamous epithelia to columnar epithelia, (D) external signals induced by acid, (E) mesenchymal to epithelial transition. Figure adapted from Phillips et al 2011 [6].

7.7 CELLULAR ORIGIN

BE is a disorder which effects the lower section of the oesophagus and causes the cells to change from normal squamous epithelial cells to columnar cells (IM) due to prolonged acid exposure. The inability to observe this process *in vivo* and a lack of animal models has meant that the true cellular origin of BE has yet to be conclusively established [62].

7.7.1 Hypotheses

Initial studies suggested the cells migrate upwards from the stomach, specifically the gastro-oesophageal junction (GEJ) [63], however this theory was later discounted when animal models suggested that the cells arose within the oesophagus itself [64-66]. Many hypotheses for the origin of the columnar cells have since been based on the latter theory. A diagrammatic overview of the hypotheses can be seen in Figure 7.2 (adapted from Phillips et al (2011) [6]). The first hypothesis was that a stem cell, located in the basal layer of the squamous epithelium, was reprogrammed to produce columnar cells rather than the normal squamous cells [67, 68]. The second hypothesis described the location of the stem cell inside the oesophageal gland duct, which

connects to the surface of the oesophagus. It was hypothesised that chronic reflux caused exposure of the stem cell which then initiated migration and differentiation of cells, from the duct, to columnar cells in order to replace the effected normal epithelium [23, 69-71]. The third hypothesis described the trans-differentiation [72] of normal squamous cells to columnar cells, caused by an acidic environment. It was hypothesised that this occurred through an epigenetic effect where the lining of the oesophagus reverts back to the cells present in early embryogenesis (columnar cells) before the trans-differentiation of these cells to mature squamous cells [73-75]. The fourth hypothesis described indirect effects on the lining of the oesophagus, specifically, changes to the epithelial layer caused by regulatory signals produced by stromal cells in the submucosa [76]. The final hypothesis suggests that the columnar cells arise from stromal cells directly via a mesenchymal-to-epithelial transition [69].

7.7.2 Gastro-Oesophageal Junction Origin

More recent studies have proposed that BE does in fact originate from the GEJ [77, 78]. In particular, a study published by Lavery et al (2014) [79], found that although segments of BE contained both intestinal (usually found in the intestine) and gastric cells (usually found in the stomach), the glands themselves were more similar to the

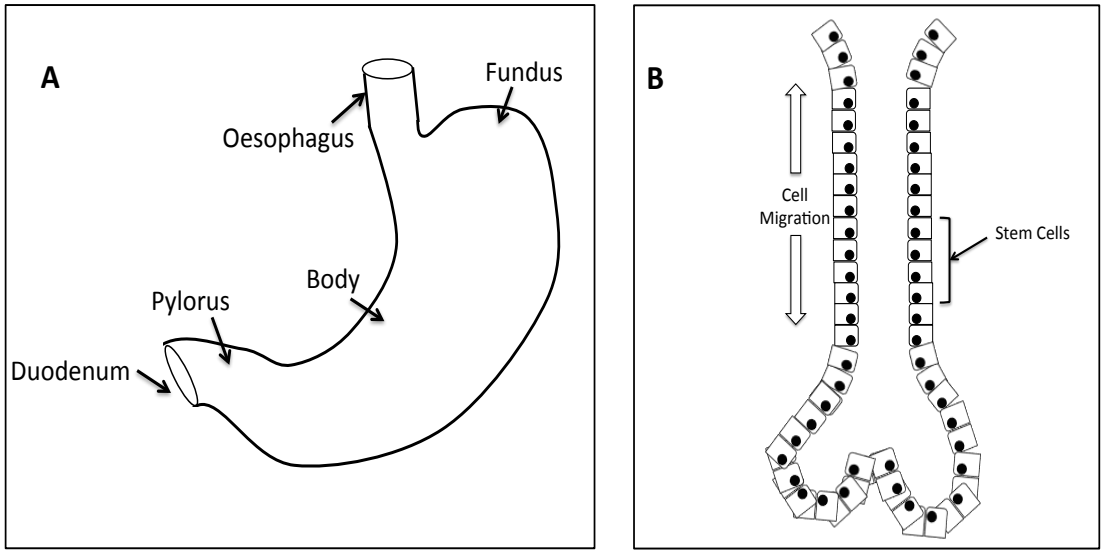


Figure 7.3: Stomach compartments and a diagrammatic view of an oesophageal gland duct. The oesophageal gland ducts (B) are similar to those found in the Pylorus of the stomach (A). The stem cells are located within the neck of the gland and cells migrate upwards towards the surface and down towards the case of the duct (bidirectional migration).

pyloric glands found at the base of the stomach (Figure 7.3A). Both pyloric and Barrett's glands have the same type of structure (Figure 7.3B). The stem cells in both glands are located in the neck, whereas in intestinal glands they are located in the base. They also found that BE glands were clonal, containing multiple multipotent stem cells, which were capable of differentiating into both intestinal and gastric cell lineages. When undertaking experiments using Ki67 and iododeoxyuridine (IdU) labelling, they found that cells moved both up to the surface and down to the base of the crypt from the neck, known as bidirectional migration (also seen in the pylorus glands).

7.8 MOLECULAR MECHANISMS

7.8.1 Pathways

Mucosal damage (e.g. reflux) is thought to result in the activation of molecular mechanisms involved in embryogenesis and/or adult tissue. Pathways thought to be involved are the Hedgehog (HH) [76, 80-84], Transforming Growth Factor β (TGFβ) [85, 86], Notch [80, 85, 86] and Mitogen-Activated Protein Kinase (MAPK) [87, 88] signalling pathways.

7.8.1.1 Hedgehog Signalling Pathway

The HH signalling pathway, of which there are three homologues: Desert (DHH), Indian (IHH), and Sonic (SHH), is important in normal embryo development. It plays an important role in the early oesophagus (columnar epithelia) but then diminishes in the mature oesophagus (squamous epithelia), resulting in the absence of the HH ligand on the epithelial surface [76, 81, 82]. Whilst this is the case in the normal oesophagus, some studies have shown that in BE, both SHH and IHH are up regulated. Wang et al (2010) [84] also found that the HH target genes *Ptch1* (*Patched 1*) and *BMP4* (*Bone Morphogenetic Protein 4)* were expressed in the stroma of BE but not of the normal epithelia. They proposed that hedgehog ligand expression could contribute to BE through the stimulation of the target genes, which in turn, triggered the production of columnar epithelia [6, 84].

7.8.1.2 Transforming Growth Factor β (TGFβ) and Notch Signalling Pathway

Both TGFβ and Notch signalling pathways are important in the developing embryo and adult tissues. The TGFβ pathway regulates multiple cellular processes, such as cell growth, differentiation and apoptosis. The Notch signalling pathway regulates cell-fate determination in the embryo and maintains homeostasis in the adult tissue [89]. A study published by Mendelson et al (2011) [86] discovered that a dysfunctional TGFβ signalling pathway was present in 5/10 BE cases and 17/22 EAC cases and that there was consistent activation of the Notch signalling pathway in EAC.

7.8.1.3 Mitogen-Activated Protein Kinase (MAPK) Signalling Pathway

The MAPK signalling pathway regulates multiple cellular processes including proliferation, differentiation and apoptosis. Studies have shown that these pathways can be activated by reflux exposure in Barrett's cell lines [87, 88]. When the cells were exposed to acid, proliferation and survival increased, whilst apoptosis decreased, suggesting that acid exposure might contribute to the metaplasia-dysplasia-carcinoma sequence seen in BE, through activation of MAPK pathways [87].

7.8.2 Chromosome Instability – Somatic Variations

Chromosome instability in BE has been illustrated in various studies [90-92]. They revealed that copy number variations (CNVs) and loss of heterozygosity (LOH) (particularly of chromosome arms 9p and 17p) increased in frequency and size between early (non-dysplastic and LGD) and late (HGD and carcinoma) stage BE (P<0.001). Li et al (2008) [90] found that chr9p had large regions of copy loss and LOH spanning most of the arm in patients with early-stage BE. Three statistically significant LOH events seen on chr9p spanned 9.0-12.1Mb, 20.5-25.0Mb and 28.5-30.5Mb. Copy loss of the *FHIT* and *WWOX* loci were also seen in patients with early-stage BE. In addition, patients with early-stage BE also had statistically significant small abnormalities, containing multiple genes, such as copy gains on 8q24.3 and 10q22.1. In contrast, patients with late-stage BE had copy loss or LOH of whole chromosome arms, such as 3p, 5p and 5q, 9p, 13p and 13q, 17p and 18q.

Similarly, Paulson et al (2009) [91] found that the most common region of CNVs in patients with early-stage BE was on chromosome 9p, specifically loss of the *p16* locus and two other areas (10.4Mb-11.8Mb and 25.5Mb-27.5Mb). Chromosomes 1 and 8 also had copy number losses. Paulson et al (2009) [91] concluded that patients with CNVs involving >70Mbp were at increased risk of progression to DNA abnormalities and/or EAC (*P*=0.0047) [91].

Many functional candidate genes for BE progression have been identified, including the highly conserved homeobox (HOX) gene family (specifically, *Caudal type homeobox (Cdx) 1* and *Cdx2*), tumour suppressor genes (p53, p63, p16 and APC (Adenomatous polyposis coli)), and other speculative genes. These genes have been implicated in small replication studies, but have not been assigned genome-wide statistical significance. More recently however, Genome-Wide Association Studies (GWAS) have identified two statistically significant BE susceptibility SNPs [2] and four statistically significant BE/EAC associated SNPs [3].

7.8.2.1 Cdx1 and Cdx2

HOX genes encode developmental transcription factors important for development [93-95]. *Cdx1* and *Cdx2*, predominantly expressed in the small intestine and colon, are believed to direct the development and differentiation of the columnar epithelium [95, 96]. Due to their role in the intestine, it was thought they might have a role in the development of BE [97]. This theory has recently been supported via a study published by Ren et al (2014) [47], which has shown that five SNPs within *Cdx1* and three within *Cdx2* were associated with BE susceptibility (*P*<0.05).

7.8.2.2 Tumour Suppressor Genes: p53, p63, p16/CDKN2A and APC

Loss of *p53* has long been established as a key step in oncogenesis, whilst mutations within exons 5-8 of this gene are seen as late steps within the MCS. *p53* mutations are seen in 5–10% of cases with unspecified dysplasia, 65% of patients with LGD, 75% of patients with HGD, and 50–90% of patients with EAC [11, 98]. On the other hand, *p63* (a member of the *p53* family of transcription factors) mutations are seen in the early

stages of BE [99, 100]. In the normal squamous epithelium, *p63* is absent. In columnar epithelium, however, *p63* is expressed in the basal layer. The importance of this gene during oesophageal development has been shown through the use of *p63* knockout mice, where they developed a columnar epithelium rather than squamous in the oesophagus [101]. *p16* (also known as cyclin-dependent kinase inhibitor 2A), is known for its importance in cell cycle regulation. This tumour suppressor protein has been shown to be inactivated in BE and EAC through the hyper-methylation or mutation of the CDKN2A promoter [102, 103]. *APC* mutations have long been established in colorectal cancer. Mutations in this gene have also been implicated in EAC, where there is increased LOH in late-stage BE [11, 104]. It has been hypothesized by Jankowski et al (1999) [11] that this may lead to reduced β-catenin degradation, increasing β-catenin levels, and facilitating epithelial-to-mesenchymal (EMT) transition via transcription factors [105].

7.8.2.3 Other BE Susceptibility Genes

Replication studies have identified candidate polymorphisms associated with BE, including: *IL-1* [106]; *Myo9B* [44]; *EGR* [107]; *TGFβ* [45]; IL-18 receptor-accessory protein and the IL-18 promoter [108]; *PXR* [109]; *IGF1*, *GHR*, *IGF1R* [43, 110]; *IL-12B* [111]; *NQO1* [112]; *GSTP1* [40]; *CCND1* [113]; *XRCC1* [114]; *IL-10* [115] and *GTSP1* [116]. SNPs near or within these genes have been associated with BE susceptibility loci in small replication studies, consisting of 22-257 cases and 94-455 controls.

7.8.3 Chromosome Instability – Germline Variations

Three germline variations were identified by Orloff et al in 2011 [46]. The three genes associated with BE/EAC (P<0.001) were; *MSR1* (Macrophage Scavenger Receptor 1), *ASCC1* (activating signal co-integrator 1 complex subunit 1), and *CTHRC1* (collagen triple-helix repeat-containing 1) [46]. *MSR1* was most frequently mutated with 8/116 patients (6.9%) in the initial study and 2/58 patients (3.4%) in the replication study.

7.8.4 Single-Nucleotide Polymorphisms identified via Genome-Wide Association Studies

Genome-Wide Association Studies examine many common genetic variants in individuals to see if any variant is associated with a trait, for example BE. The first GWAS (comprising 7,838 cases and 17,997 controls) published in 2012 by Su et al (2012) [2] identified two SNPs, one near *FOXF1* and another within the MHC region, to be associated with BE [2]. The *FOXF1* gene encodes the transcription factor FOXF1 (Forkhead Box F1) protein. FOXF1 is important in the development of pulmonary mesenchyme (the embryonic tissue from which blood vessels of the lung arise) and the development of the GI tract. The gene-rich MHC region is important for immune system regulation. Defects within MHC genes can result in autoimmune disorders, such as multiple sclerosis and inflammatory bowel disease [117, 118].

A second GWAS (comprising 2,363 EAC cases, 3,116 BE cases and 10,060 controls) identified four SNPs (three loci) to be associated with BE and EAC (combined) located within *CRTC1* and *BARX1* and near *FOXP1* [3]. *CRTC1* encodes the CREB-regulated transcription coactivator 1 (CRTC1) protein. *BARX1* encodes a member of the Bar subclass of homeobox transcription factors, BARX homeobox 1 (BARX1) protein. BARX1 is thought to regulate differentiation of stomach epithelia via the WNT signalling pathway [119, 120]. *FOXP1*, a tumour suppressor gene, encodes a member of the FOX transcription factors, FOXP1. This particular transcriptional repressor is thought to be important in the specification and differentiation of the lung [121-123].

7.9 CLINICAL TRIALS

Clinical trials provide important information and samples for disease research. There are currently two on-going UK-based trials for Oesophagitis, BE and EAC: ChOPIN and AspECT.

7.9.1 Chemoprevention Of Premalignant Intestinal Neoplasia (ChOPIN)

The ChOPIN study aims to assess both epigenetic and genetic changes that lead to pre-malignancy and cancer. The study has been devised in two parts; firstly,

biomarkers, array and assessment of clonal changes in pre-malignant mucosa and early cancer. Secondly, a GWAS to investigate inherited predisposition to oesophageal diseases. Each patient in the study is required to complete a patient history form which records information on height, weight, age, diet, smoking history, alcohol consumption and surgical procedures; all of which will be useful for performing sub-analyses.

7.9.2 Aspirin and Esomeprazole Chemoprevention in Barrett's Metaplasia (AspECT)

The AspECT study is investigating the use of both aspirin and PPI therapy. The study aims to investigate the benefits of acid suppression with low or high dose PPI (esomeprazole) with or without aspirin in reducing the risk of cancer and/or HGD in BE patients. It is also examining whether intervention with aspirin with or without PPI therapy results in decreased mortality rate.

7.10 BARRETT'S OESOPHAGUS GENOME-WIDE ASSOCIATION STUDY

The research completed for this thesis is a continuation of a previously published GWAS paper by Su et al (2012) [2]. The sample sets and how they were used in the previous paper is outlined below. Sample sets described below have been used in the replication study (current research) when undertaking meta-analyses. All phases consist of unrelated individuals of white Caucasian origin and each sample set is independent of one another.

7.10.1 Samples

7.10.1.1 Cases

All cases were diagnosed with histologically-confirmed BE, accompanied with an endoscopy report. Cases in the Discovery Phase were recruited under the UK-based AspECT clinical trial [124]. Sample collection was in accordance with the British Society of Gastoenterology criteria [53], which is the standard practice for histopathologists in the UK and most of Europe. In the Discovery set, 90% of the samples had evidence of IM and so also met the American College of

Gastroenterology criteria used in the United States of America [57].

The UK, Irish and Dutch Replication Phases were acquired from the ChOPIN genetic study and the Esophageal Adenocarcinoma GenEtics (EAGLE) consortium [7].

Discovery cases

1,852 BE UK cases (80.3% male) from multiple NHS sites across the UK, collected under AspECT (Chief Investigator: J Jankowski).

Stage 1: UK Replication 1 (UKREP1) cases

1,105 BE patients (70.9% male, 0.3% not stated) from multiple NHS sites across the UK recruited under ChOPIN.

Stage 2: Dutch Replication cases

473 BE patients (74.0% male) were collected under ChOPIN from the University Medical Centre, Groningen.

Stage 3: UK Replication 2 (UKREP2) cases

1,765 BE cases (71.4% male, 0.6% not stated) collected from various NHS sites across the UK under ChOPIN.

Stage 3: Irish Replication cases

245 BE cases (64.1% male, 8.6% not stated) from St. James's Hospital and Mater Misericordiae University Hospital, Dublin.

Stage 3: BEACON Replication cases

2398 cases (76.0% male) were collected as part of a GWAS study (BEAGESS). Samples were collected from sites in Australia (n=325), Europe (n=363) and North America (n=1710).

7.10.1.2 Controls

Controls for each phase were collected from various sources detailed below.

Discovery Controls

5,172 population controls, part of the Wellcome Trust Case Control Consortium 2 (WTCCC2; Chief Investigator: P Donnelly) set (50.4% male), made up of the 1958 British Birth Cohort (58C) and the National Blood Service collections (UKBS).

A total of 2,673 controls were selected from the 58C (also known as The National Child Development Study (NCDS)). The cohort was set up to follow the lives of 17,000 people born in England, Scotland and Wales in a single week in March 1958. Each of the surviving cases were followed up at ages 7, 11, 16, 23, 33, 42, 46, 50 and 55 [125]. The remaining 2,499 controls consisted of blood donors recruited by the WTCCC in collaboration with the UK Blood Services [126].

Stage 1: UKREP1 controls

A total of 6,819 controls (47.9% male) were used for UK Replication 1. 2,578 controls were part of People of the British Isles (PoBI; Chief Investigator: W Bodmer) and the remaining 4,241 controls were samples from the 58C that had not been used in the Discovery Phase.

PoBI, funded by the Wellcome Trust, began collecting blood samples from 4,500 people from rural populations throughout the British Isles in 2004 to examine genetic differences around the UK. A sub-section of the study is also investigating the inherited variation of facial features within the UK.

Stage 2: Dutch Replication controls

1,780 controls (59.2% male) provided by the University Medical Centre, Groningen.

Stage 3: UKREP2 controls

1,586 controls (45.2% male, 4.3% not stated) were collected under the Colorectal Tumour Gene Identification (CoRGI) Consortium (Chief Investigator: I Tomlinson). CORGI aims to identify genes that increase the risk of bowel cancer or non-cancerous tumours. The controls used in this study consist of individuals unaffected by cancer and without a family history of colorectal neoplasia [127].

Stage 3: Irish Replication controls

A total of 473 controls (35.1% male, 0.6% not stated) were provided by Trinity Biobank, Dublin. The Biobank was established in 2004 and facilitates processing, storage and distribution of specimens for research undertaken in Trinity College (St James's Hospital).

Stage 3: BEACON Replication controls

A total sum of 2167 controls (78.6% male) were collected as part of a GWAS (BEAGESS). Samples were collected from sites in Australia (n=561), Europe (n=333) and North America (n=1273).

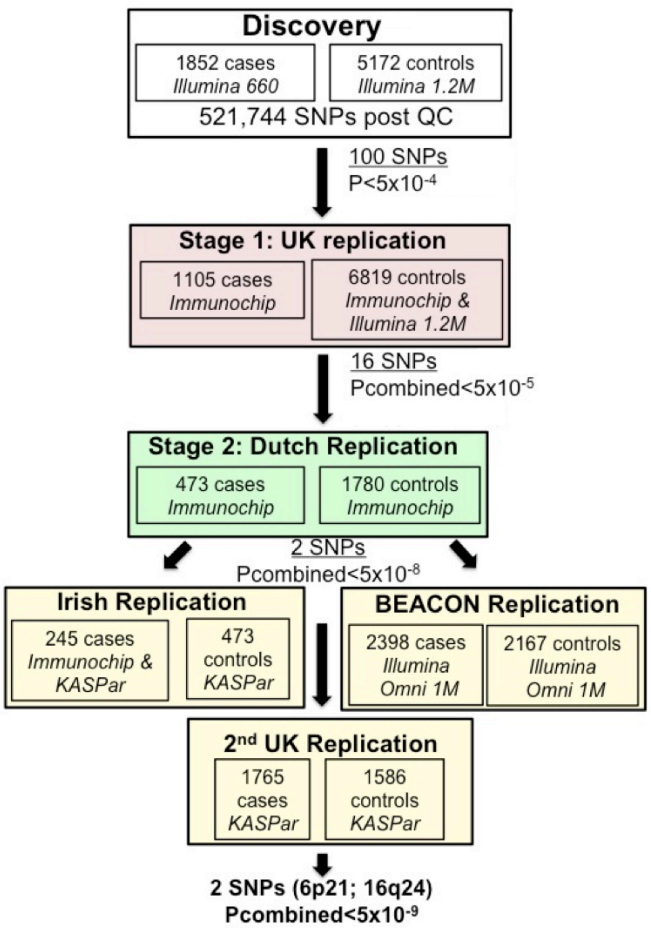


Figure 7.4: Stages of the previous Genome-Wide Association Study paper. Figure adapted from Su et al [2]

7.10.2 SNP Selection

In the Discovery Phase 521,744 SNPs were analysed. After analysis of the Discovery SNPs, only 100 with $P<5\times10^{-4}$ were taken forward to Stage 1 Replication. When combining the Stage 1 Replication data with the Discovery data, there were 16 SNPs with $P<5\times10^{-5}$. These 16 SNPs were taken forward to Stage 2 Replication. Once the Stage 2 data had been combined with Discovery and Stage 1, two SNPs reached genome-wide significance, with $P<5\times10^{-8}$. These 2 SNPs were further replicated in Stage 3 for validation, where both SNPs reached $P<5\times10^{-9}$ and were reported as BE susceptibility loci. An overview of sample sets and SNPs analysed can be seen in Figure 7.4 (adapted from Su et al (2012) [2]).

7.10.3 Ethical Considerations

Written informed consent was obtained from all subjects. The ethics of the project were reviewed by the East London and the City Research Ethics Committee (04/Q0603/1). All UK studies were implemented with national ethics committee approval (MREC numbers: AspECT 04/Q0603/1; ChOPIN/IPOD 06/Q1603/07; HANDEL 09/H0505/23; and CORGI 06/Q1702/99). The Irish samples were collected with approval from the Research Ethics Committee Board of St. James's Hospital. The Dutch replication samples were collected with approval from the ethics committee or institutional review board of all participating institutions. The BEACON/BEAGESS project obtained informed consent from all recruited participants, and was approved by the ethics boards of each participating institution.

7.10.4 Genotyping

7.10.4.1 Discovery Phase

Genotyping was performed using the Illumina 660W-Quad array for cases and a custom Human 1.2M-Duo array for controls at the Wellcome Trust Sanger Institute (WTSI) [2].

7.10.4.2 Stage 1

UKREP1 genotyping was performed using the custom Illumina Infinium HD genotyping array, the Immunochip, at the WTSI [2].

7.10.4.3 Stage 2

Dutch Replication samples were genotyped on the Illumina Immunochip but in two separate locations; the cases were genotyped at WTSI and the control samples were previously genotyped [128].

7.10.4.4 Stage 3

(a) Irish Replication: 168 cases were genotyped on the Illumina Immunochip at WTSI. rs9257809 and rs9936833 (the 2 SNPs taken forward to stage 3) were genotyped in 77 cases and all controls using competitive allele-specific PCR KASPar

chemistry (LGC Ltd, Hertfordshire, UK).

(b) UKREP2: All samples were genotyped using KASPar competitive allele-specific PCR.

(c) BEACON Replication: All samples were genotyped at the Fred Hutchinson Cancer Research Center (FHCRC) on the Illumina Omni1M Quad.

7.11 INDIVIDUAL CONTRIBUTIONS FOR THE REPLICATION STUDY

The Replication Study completed as part of this thesis is a continuation of the BE GWAS data published by Su et al (2012) [2].

Since joining the BE research team (who published the GWAS in 2012) in February 2013, four new sample cohorts have been collected and analysed for a specific set of SNPs.

The research completed for both the BE GWAS and the Replication study was a collaborative project between researchers both nationwide and worldwide. Table 7.1 shows the contributions of each individual involved in the BE Replication study.

Whilst completing the BE Replication Study a paper analysing 87 SNPs, identifying four as novel BE/EAC-associated SNPs, was published [3]. It was decided to try to replicate the 87 SNPs published by Levine et al (2013) [3]. Table 7.2 shows the individual contributions for the Levine replication study.

Section	Researcher	Comments
Collection and Extraction of Replication Phase 2 case and control DNA	Available from previous publication [1]	
Replication Phase 2 SNP selection	Claire Palles, Janusz Jankowski, Ian Tomlinson	
Sequenom iPLEX design	Claire Palles	
Sequenom iPLEX genotyping	Wellcome Trust Sanger Institute	
Replication Phase 3 SNP selection	Claire Palles, Janusz Jankowski, Ian Tomlinson	
Processing of ChOPIN/AspECT samples for screening	Laura Chegwidden, Barbara Zietek, Neera Maroo, Laura Gay, Manoj Nanji	Time Periods: Laura Chegwidden: Feb'13-Feb'14; Barbara Zietek: Jan'12-Apr'13; Neera Maroo: Jan'12-Dec'12; Laura Gay: '08-'12; Manoj Nanji: '06-'09
UK Replication 3 case collection, processing and storage	Laura Chegwidden, Barbara Zietek, Neera Maroo, Laura Gay, Manoj Nanji	610 cases by Laura Chegwidden; 387 cases by Barbara Zietek, Neera Maroo, Laura Gay and Manoj Nanji
Phenotype Classification and Patient History Form input for UK Replication 3.	Laura Chegwidden and Laura Gay	BE and IM Classification for 610 cases by Laura Chegwidden; 387 cases by Laura Gay
UK Replication 3 case DNA extraction	Laura Chegwidden, Claire Palles and John Findlay	610 cases by Laura Chegwidden; 387 by Claire Palles and John Findlay
UK Replication 3 control DNA extraction	Elinor Sawyer and Rebecca Roylance	
UK Replication 3 case genotyping	Laura Chegwidden, Claire Palles and John Findlay	610 cases by Laura Chegwidden; 387 by Claire Palles and John Findlay
UK Replication 3 control genotyping	Claire Palles and Sarah Briggs	4 SNPs by Claire Palles, 3 SNPs by Sarah Briggs
Belgian case DNA extraction	Hans Prenen	
Belgian control DNA extraction	Isabelle Cleynen	
Belgian case genotyping	Claire Palles	
Belgian control genotyping	Claire Palles	
Dutch Extension case DNA extraction	Auke Verhaar	
Dutch Extension control DNA extraction	Kausila Krishnadath	
Dutch Extension case genotyping	Claire Palles	
Dutch Extension control genotyping	Claire Palles	
BEACON case/ control genotypes for the 7 SNPs selected for Replication Phase 3.	BEACON members	Genotypes supplied by BEACON/BEAGESS
BE/EAC SNP analysis	Claire Palles and Laura Chegwidden	
Imputation	Claire Palles	
Functional Analysis	Laura Chegwidden and Claire Palles	
Association Testing and Meta-analysis	Laura Chegwidden, Claire Palles and Xinzhong Li	

Table 7.1: Contributions of individuals at each stage of the BE Replication study.

Section	Researcher	Comments
Levine SNP selection	Laura Chegwidden, Claire Palles, Ian Tomlinson, Janusz Jankowski	
Imputation of all 87 Levine SNPs	Claire Palles	
Genotyping 4 SNPs selected from the remaining 83 Levine SNPs	Laura Chegwidden, Claire Palles, Claire Adams, John Findlay	Laura Chegwidden: genotyped 4 SNPs in Replication Phases 1, 2 and 3, genotyped 1 SNP in Discovery; Claire Palles: Aided genotyping when available; Claire Adams: Aided case DNA plating; John Findlay: Helped genotype 1 SNP
Principle component analysis	Claire Palles	
Association Testing and Meta-analysis	Claire Palles, Laura Chegwidden, Xinzhong Li	

Table 7.2: Contributions of individuals at each stage of the Levine study replication.

8. MATERIALS AND METHODS

The research carried out for this degree is a replication study of the BE GWAS paper published by Su et al (2012) [2]. There is an overlap of samples used in UKREP2 and the Irish Replication. However, the SNP selection between the BE GWAS paper [2] and the replication study is different, due to increasing the *P* value threshold after combining data from Discovery, UKREP1 and the Dutch Replication sample sets ($P<10^{-4}$ rather than $P<10^{-8}$; described in section 8.4.1).

In this section, a description of the cases and controls, genotyping methods, SNP selection, quality control and statistical methods used for the replication study will be provided.

8.1 SAMPLE SETS

Cases and controls used in each phase of the replication study are detailed below. An overview of the BE GWAS samples and the Replication study samples is provided in Table 8.1. All phases consist of unrelated individuals of white Caucasian origin and each sample set is independent of one another.

8.1.1 Cases

BE cases within the UK, Irish, Belgian and Dutch Replication Phases were acquired through the ChOPIN study and the EAGLE consortium [7].

Patient recruitment to ChOPIN consisted of a baseline blood sample and completion of a Patient History Form (shown in the appendices: Figure 11.1), to record all phenotypic information (such as dysplasia and presence/absence of IM). All cases were initially diagnosed with BE by endoscopy, which was subsequently confirmed by histology. Individuals with BE lengths of ≥1cm (C1M1) circumferential disease or ≥2cm tongue patterns (C0M2), according to the Prague criteria [129] (Figure 8.1, from www.iwgco.net), were recruited to ChOPIN. Participants included in this research were unrelated and of white Caucasian origin. Presence of EAC in these patients

		BE GWAS [1]					Replication Study			
		Discovery	Replication Phase 1		Replication Phase 2		Replication Phase 3			
			UKREP1	Dutch Replication	Irish Replication	UKREP2	UKREP3	Belgian Replication	Dutch extension	BEACON/ BEAGESS
	N Cases (IM+)	1852 (1667)	1105 (729)	473 (378)	245 (245)	1765 (1465)	997 (603)	341 (341)	64 (4)	3295 (3295)
	N Controls	5172	6819	1780	473	1586	974	848	206	3204
Genotyping Method	Case	Illumina660	Immunochip	Immunochip	iPLEX & Immunochip	iPLEX	KASPar	Immunochip & KASPar	KASPar	Illumina Omni 1M
	Control	Illumina1.2M	Immunochip & Illumina1.2M	Immunochip	iPLEX	iPLEX & Illumina Hap550	KASPar	Immunochip	KASPar	Illumina Omni 1M

Table 8.1: Numbers of cases and controls, genotyping methods used and number of cases with intestinal metaplasia (IM+). Note that Replication Phase 2 samples were also included in the BE GWAS [1], but only for two SNPs that were genotyped by KASPar, not the 65 SNPs analysed in this Replication Study. BE, Barrett's Oesophagus; GWAS, Genome-wide association study; N, Number.

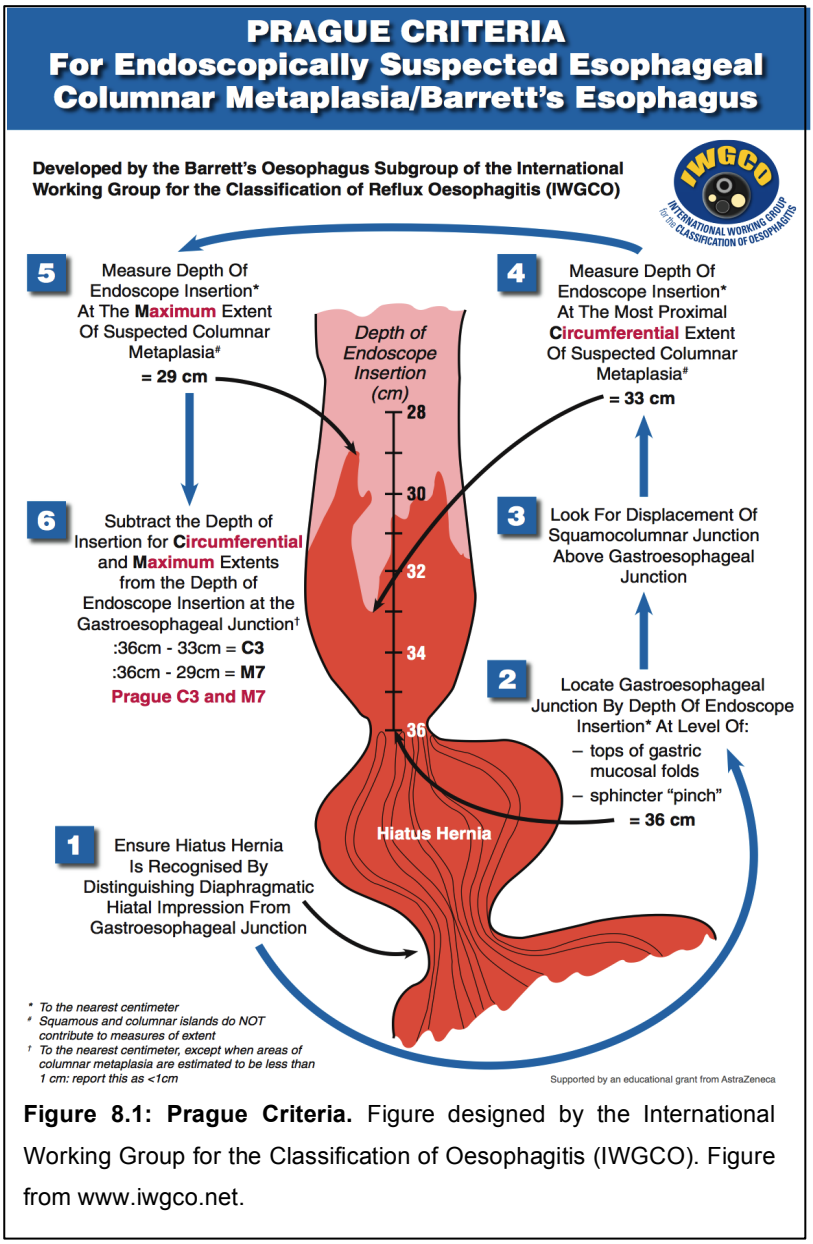


Figure 8.1: Prague Criteria. Figure designed by the International Working Group for the Classification of Oesophagitis (IWGCO). Figure from www.iwgco.net.

(either at the time of recruitment or afterwards) was recorded but was not an inclusion/exclusion criterion.

8.1.1.1 Replication Phase 2 cases

UKREP 2 cases

1,765 BE cases (71.4% male, 0.6% not stated) from NHS sites across the UK.

Irish Replication cases

245 BE cases (64.1% male, 8.6% not stated) from St. James's Hospital and Mater Misericordiae University Hospital, Dublin.

8.1.1.2 Replication Phase 3 cases

UK Replication 3 (UKREP3) cases

997 BE cases (70.9% male, 1.6% not stated) from NHS sites across the UK.

Belgian Replication cases

362 BE cases (66.0% male, 12.3% not stated) from Leuven, Belgium.

Dutch Extension cases

64 cases (28.1% male, 68.8% not stated) from Nijmegen and Rotterdam, Netherlands.

BEACON cases

A total of 3,295 BE cases (75.5% male), predominantly of Northern European descent, collected as part of the BEACON consortium GWAS (Chief Investigators: T Vaughan, D Whiteman, D Levine).

8.1.2 Controls

Control subjects were derived from various sources outlined below.

8.1.2.1 Replication Phase 2 controls

UKREP2 controls

1,586 controls (45.2% male, 4.3% not stated) from the Colorectal Tumour Gene Identification (CoRGI) Consortium (Chief Investigator: I Tomlinson). This study aims to identify genes that increase the risk of bowel cancer or non-cancerous tumours (polyps and adenomas). The controls used in this study consisted of individuals unaffected by cancer and without a family history of colorectal neoplasia [127].

Irish Replication controls

473 controls (35.1% male, 0.6% not stated) were provided by Trinity Biobank, Dublin. The Biobank, established in 2004, facilitates processing, storage and distribution of specimens for clinical and epidemiological research undertaken in Trinity College (St James's Hospital).

8.1.2.2 Replication Phase 3 controls

UKREP3 controls

974 female controls from the Genetics of LobulAr Carcinoma In situ in EuRope (GLACIER) study (Chief Investigators: E Sawyer, R Roylance). GLACIER aims to identify genetic changes in patients with lobular carcinoma, in situ, of the breast to gain a greater depth of understanding of the disease and possible treatment options. The controls used in this study consist of individuals with no personal or family history of breast cancer [130].

Belgian Replication controls

848 controls (47.6% male) from Leuven, Belgium.

Dutch Extension controls

206 controls (61.7% male) from Nijmegen and Rotterdam.

BEACON controls

3,204 controls (72.6% male), predominantly of Northern European descent, were provided by the BEACON consortium GWAS (BEAGESS).

8.2 ETHICAL CONSIDERATIONS

Written informed consent was obtained from all subjects. The ethics of the project were reviewed by the East London and the City Research Ethics Committee (04/Q0603/1). All UK studies were implemented with national ethics committee approval (MREC numbers: ChOPIN/IPOD 06/Q1603/07; CORGI 06/Q1702/99; and GLACIER 06/Q1702/64). The Irish samples were collected with approval from the Research Ethics Committee Board of St. James's Hospital. The Dutch extension and Belgian replication samples were collected with approval from the ethics committee or institutional review board of all participating institutions. The BEACON/BEAGESS project obtained informed consent from all recruited participants, and was approved by the ethics boards of each participating institution.

8.3 GENOTYPING

8.3.1 DNA Extraction, Quantification and Concentration

Genomic DNA (gDNA) for Replication Phase 2 (UKREP2 and Irish) cases and controls had been extracted and quantified previously for the published BE GWAS [2].

Case and control gDNA for UKREP3 and the Dutch Extension were extracted using Maxwell® 16 Research Instrument (Promega UK) at the Wellcome Trust Centre for Human Genetics, Oxford (WTCHG).

For UKREP3 samples, gDNA was quantified using the Nanodrop 2000 (Thermo Fisher Scientific Inc) UV-Vis Spectrophotometer. Concentrations recorded by the Nanodrop instrument were halved, to allow for over-estimation of gDNA content. Samples with a concentration of >20ng/µl and an A260/280 score of 1.5-2.5 were deemed sufficient for genotyping (N=302). Samples with a concentration <20ng/µl or with an A260/280 score of <1.5 or >2.5 were concentrated using SPD2010 Integrated SpeedVac™ Systems (Thermo Fisher Scientific Inc) and re-suspended in molecular grade water to produce a concentration of 20ng/µl (N=308). Nanodrop readings were re-taken for the 308 gDNAs, samples which still failed to meet the criteria were re-extracted (N=16).

Belgian case and control gDNAs were extracted and quantified by Hans Prenen and Isabelle Cleynen, respectively, in Belgium.

Dutch Extension case and control gDNAs were extracted and quantified by Auke Verhaar and Kausilia Krishnadath respectively, in the Netherlands.

8.3.2 Genotyping Methods used

8.3.2.1 Sequenom iPLEX® MassArray®

Three custom Sequenom iPLEX® MassArray® platforms were designed by Dr Claire Palles and genotyped at the WTSI. The assay consists of multiple steps [131]. First, the gDNA undergoes a locus-specific PCR reaction. Secondly, all unincoporated dNTPs are removed using shrimp alkaline phosphatase (SAP), which cleaves a phosphate from the unincorporated dNTPs, converting them to dNDPs and rendering them unavailable to future reaction. This is then followed by the iPLEX reaction, where there is locus-specific extension of the annealed primer upstream of the target SNP. During the iPLEX reaction, the primer and DNA are incubated with mass-modified dideoxynucleotide terminators [4, 131]. The dideoxynucleotides will have differing mass, dependent on the base incorporated, which can be assessed using Matrix Assisted Laser Desorption/Ionization-Time of Flight (MALDI-TOF) mass spectrometry. This mass difference allows the data analysis software to differentiate between SNP alleles [132]. Genotypes were assigned using MassArray® Analyzer 4 System® [132]. An overview of the Sequenom iPLEX® MassArray® workflow is outlined in Figure 8.2 (adapted from Gabriel et al (2009) [4]).

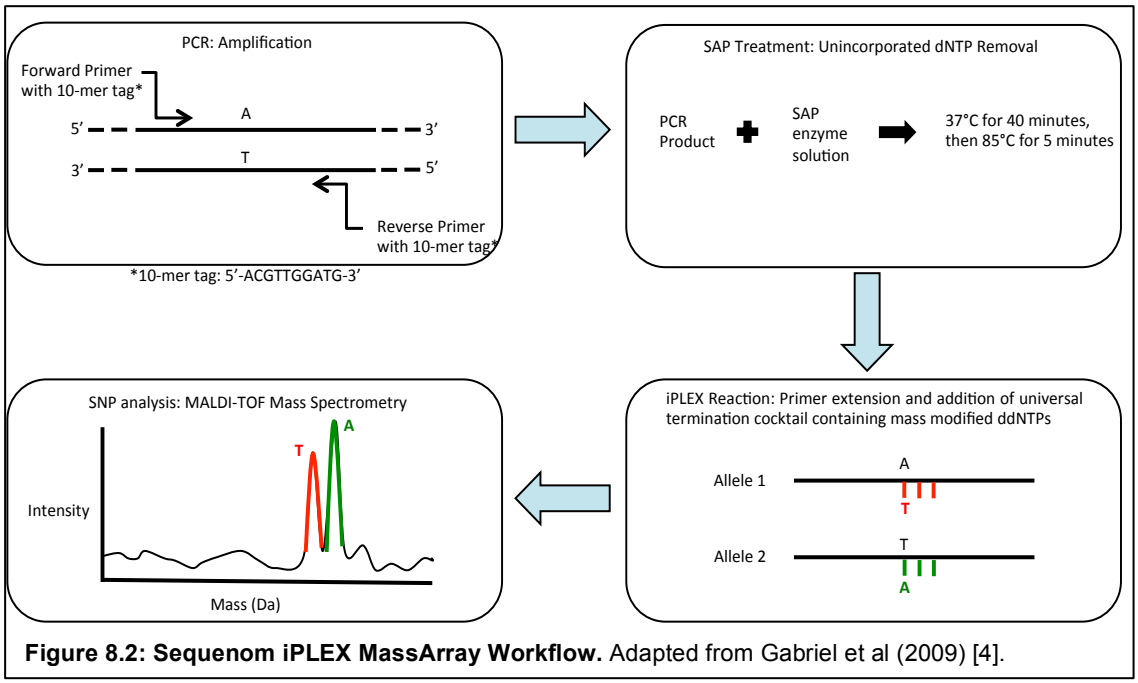


Figure 8.2: Sequenom iPLEX MassArray Workflow. Adapted from Gabriel et al (2009) [4].

8.3.2.2 Illumina® HumanHap550

The Illumina® HumanHap550 BeadChip (Illumina Hap550) contains probes for 550,352 SNPs (www.illumina.com) and utilises the Infinium® HD Assay [133]. For each SNP, specific 50-mer oligonucleotide probes (oligos) are covalently linked to beads, which are then dispersed into the wells. The Infinium® HD Assay was designed to provide high-quality, accurate results quickly. Firstly, gDNA samples are denatured and then isothermally amplified (no PCR is needed). The amplified gDNA is then fragmented by a controlled enzymatic step, producing a target gDNA fragment size of around 300bp. After an isopropanol precipitation, the fragmented gDNA is collected by centrifugation and then re-suspended in hybridization buffer. Samples are then loaded onto the BeadChip and incubated. The gDNA will anneal to the specific oligos present on the beads. gDNA which is either unhybridized or non-specifically hybridized after the previous annealing step is then washed away. Detectable labels are then incorporated on the BeadChip by Single-base extension (SBE) of the oligos, which determines the genotype call. The BeadChip is then scanned, using a laser to excite the fluorophore attached to the SBE product on the beads, and records images of the emitted light [133-136]. An overview of the workflow can be seen in Figure 8.3.

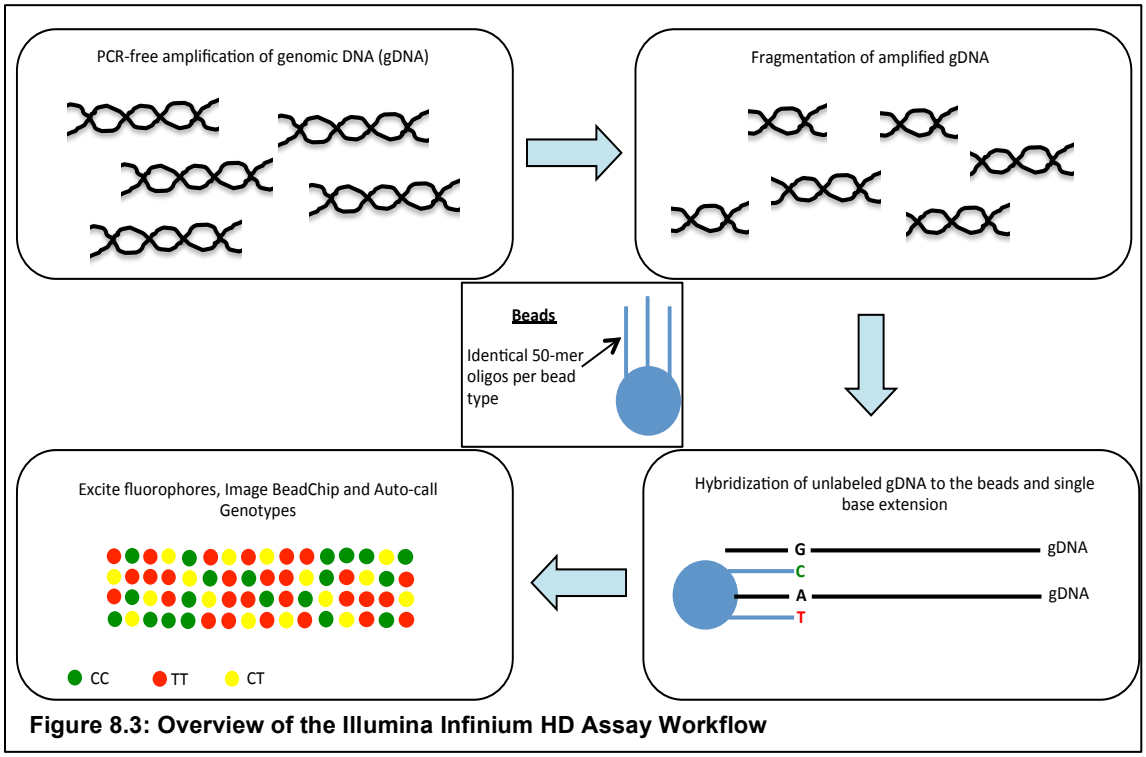


Figure 8.3: Overview of the Illumina Infinium HD Assay Workflow

8.3.2.3 Illumina® Immunochip®

The Immunochip®, initiated by the WTCCC, is an Illumina® Infinium® iSelect HD custom genotyping array, designed by the Immunochip Consortium. The chip contains 196,524 polymorphisms, located in regions of the genome that show evidence of association to immune-mediated diseases, including BE. It is designed for use in white European populations and is a relatively low-cost option compared to other GWAS chips [137]. The Illumina Infinium HD assay is described in the section above (8.3.2.2) and an overview of the workflow can be seen in Figure 8.3.

8.3.2.4 KASPar

KASPar is a homogeneous, endpoint PCR genotyping technology. The KASP™ assay relies on Fluorescence Resonance Energy Transfer (FRET) to distinguish between SNP alleles. The allele-specific primers each have a sequence that corresponds to one of two FRET cassettes (one labelled with FAM™ dye and one with HEX™ dye) on the 5' end. Allele discrimination is achieved through the competitive binding of the two allele-specific primers [138, 139].

The KASP genotyping reaction is made up of three elements; gDNA, assay mix and master mix. The assay mix for this technology consists of two allele-specific primers and one common primer. Primers were ordered desalted on the 0.025M scale and are listed in Table 11.2. The primers were made up to a concentration of 0.1nmol/µL. Firstly, 1mL of assay mix was prepared consisting of 12µL the Allele 1 primer, 12µL of the Allele 2 primer, 30µL of the common primer and 46µL of molecular grade water. The concentrations in the prepared 1mL of assay mix were as follows: 0.012nmol/µL of allele 1 primer, 0.012nmol/µL of allele 2 primer and 0.030nmol/µL of the common primer. The master mix was then prepared; each well (reaction) required 5µL KASPar 2X reaction buffer, 0.14µL of the previously prepared assay mix and 3.86µL of molecular grade water. The primer concentrations per reaction were as follows: 0.114nmol/µL of allele 1 primer, 0.114nmol/µL of allele 2 primer and 0.286nmol/µL of the common primer.

Assay Mix	Amount added (µL)
Allele 1 primer (100 µM)	12
Allele 2 primer (100 µM)	12
Common primer (100µM)	30
Water	46

Master Mix	Amount added per well (µL)
KASPar 2X Reaction Buffer	5
Assay Mix	0.14
Water	3.86

Table 8.2: KASP Assay and Master mixes. Primers must be made up to 100µM. Assay mix detailed above makes 1ml, enough for roughly 7 96-well plates. Master mix above shows the amount needed per well, to be added to the 2µL of DNA in each well.

The 2X KASPar Reaction buffer contained the FAM and HEX specific dyes, passive reference dye ROX, modified Taq polymerase for allele-specific PCR and buffer [139, 140]. The master mix was prepared in batches (enough for at least one plate at a time) on the day of use. An overview of the assay mixes can be found in Table 8.2.

To each well, of the ABI Fast Taqman plates, the following were added: 1.5µL of 20ng/µL gDNA (the minimum recommended volume and concentration of gDNA) and 9µL of the prepared Master mix (recommended minimum total aliquot per well is 10µL due to evaporation, according to LGC Genomics [140]). The plates were then sealed with ABI optical plate seals and pulse-centrifuged for 10 seconds.

The sealed plates were placed on the G-STORM^{TM} GS4 Multi Block Thermal Cycler, with heat mats to prevent excess evaporation. The gDNA then underwent a 3-step KASPar PCR reaction. The first step was activation at 94°C for 15 minutes. The second step was 10 cycles of denaturation at 94°C for 20 seconds and annealing/elongation at a 61°C-55°C temperature titration for 1 minute, dropping 0.6°C per cycle. The final step consisted of 26 cycles of denaturation at 94°C for 20 seconds and annealing/elongation at 55°C for 1 minute [140]. An overview of the PCR reaction can be found in Table 8.3.

Step	Description	Temp (°C)	Time	N Cycles
1	Activation	94	15 mins	1
2	Denature	94	20 ses	10
	Annealing/ Elongation	61-55	60 secs (drop 0.6°C per cycle)	
3	Denature	94	20 ses	26
	Annealing/ Elongation	55	60 secs	

Table 8.3: KASPar Genotyping PCR details

All plates were removed from the thermocycler at the end of the program and pulse-centrifuged for 10 seconds. The plates were then read on Applied Biosystems™ 7900HT Fast Real-Time PCR System using SDS software 2.4 (Applied Biosystems™). Plates where the genotypes were not adequately clustered were subjected to 10 extra cycles, under the temperatures seen in Step 3 of Table 8.3 [140].

An overview of the KASP genotyping technology is available in Figure 8.4.

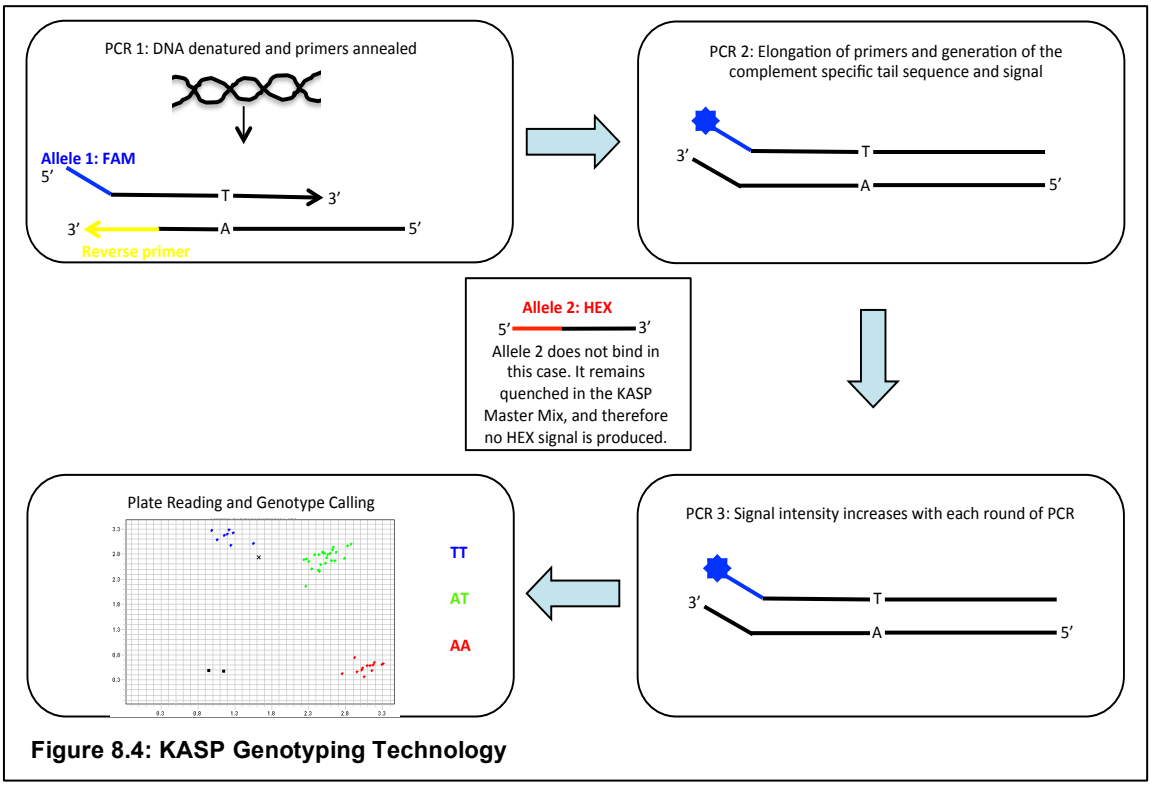


Figure 8.4: KASP Genotyping Technology

In detail, the first round of PCR entails DNA denaturation, common primer hybridization and elongation, and allele-specific primer (specific to the SNP in the assay) hybridization and elongation (with a 5' tail). During the second PCR reaction ("PCR 2" in Figure 8.4), the common primer binds to the template and extends. This produces a complementary sequence to the 5' tail created in "PCR 1". This then allows the oligos with the attached fluorophore (specific to the SNP) to hybridize. Upon hybridization the fluorophore is released from its quencher. This process continues (fluorophore hybridization and quencher release) during the amplification of the PCR product in order to obtain a strong signal [141].

Before genotyping all gDNAs, each SNP was subjected to an optimisation step. Two common primers were ordered and genotyped on one plate (half with common primer 1 and half with common primer 2). Using Applied Biosystems™ 7900HT Fast Real-Time PCR System and SDS software 2.4 (Applied Biosystems™), each half of the plate was visualised and the primer that gained tighter clustering was selected for genotyping all gDNA. An example of the optimisation step can be seen in Figure 8.5.

8.3.2.5 Illumina Omni 1M

The Illumina Omni 1M platform covers over 1 million SNPs across the genome and uses the Illumina Infinium HD assay as described in section 8.3.2.2 and seen in Figure 8.3.

All BEACON cases and controls were genotyped on the Illumina Omni 1M and genotypes were provided by the consortium.

8.3.3 Quality Control

8.3.3.1 Sequenom

Any gender discrepancies between samples or SNP call rates of <95% were excluded from the analysis.

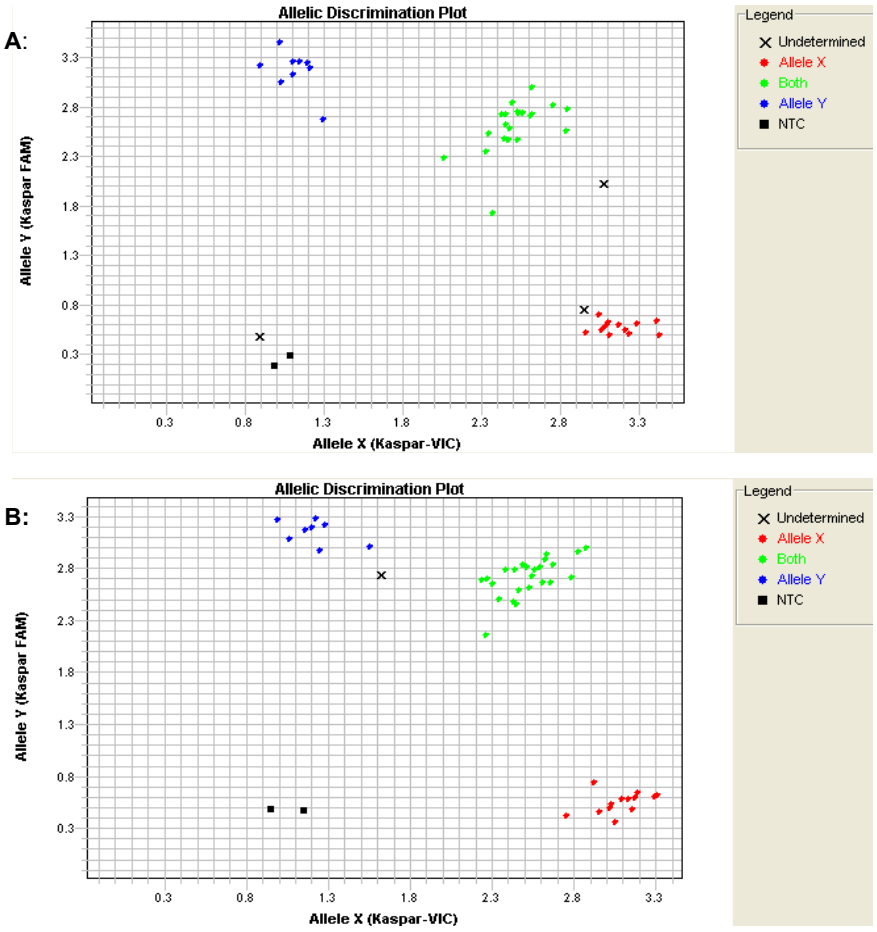


Figure 8.5: Optimisation of primers used for KASP genotyping. Showing common primer 1 (A) and common primer 2 (B). Common Primer 2 was selected in the case illustrated due to tighter clustering of the heterozygotes and high auto-call rates.

8.3.3.2 KASPar

The 7 SNPs genotyped by KASPar in Replication Phase 3 had call rates >95%. Samples were excluded if there were any gender discrepancies and SNPs were excluded if call rates were <95%. At least two negative controls were included on each plate and a positive control was used when validating each SNP and when SNPs had a low allele frequency. Genotyping quality control (QC) was tested for samples analysed solely by KASPar using duplicate DNA samples within studies and SNP assays. For all SNPs, >98% concordant results were obtained.

8.4 REPLICATION STUDY SNP SELECTION

8.4.1 Replication Phase 2 SNPs

Although the Replication Phase 2 samples have been used previously in the original BE GWAS [2], the SNP selection criteria were relaxed for the replication study. In the BE GWAS only two SNPs were genotyped by KASPar. In this replication study, 83 SNPs were selected for genotyping on Sequenom iPLEX® MassArray® provided they met one of the four criteria below, based on data (Discovery and/or Replication Phase 1) from the previous BE GWAS [2] or on candidate SNPs [46, 110].

(i) *P* value $<10^{-4}$ in combined Discovery and Replication Phase 1 analysis (N=63) from the previous BE GWAS [2]

(ii) *P* value $<10^{-4}$ in Discovery Phase, but not included in Replication Phase 1 (N=12) published by Su et al (2012) [2]

(iii) *P* value $<10^{-4}$ in a sex-stratified analysis of the Discovery phase (N=5) in Su et al (2012) [2]

(iv) Candidate polymorphisms previously reported as associated with BE and not well tagged by the Discovery Phase or Replication Phase 1, specifically, *MSR1* p.Arg293Gly [46], and variants in *IGF1R* and *GHR* [110] (N=3).

Sequenom iPLEX assays (as described in section 8.3.2.1) were successfully designed for 65 of the SNPs outlined above (Table 11.1). Any SNP in the top 40, ranked by *P* value, which failed during the design stage of the Sequenom iPLEX, was genotyped by KASPar (N=3). Unfortunately, 18 SNPs on the iPLEX design could not be analysed in the Irish replication set, as they were not present on the Immunochip (described in section 8.3.2.3).

8.4.2 Replication Phase 3 SNPs

SNPs selected for genotyping in Replication Phase 3 samples were based on a meta-analysis between Discovery set and Replication Phase 1 samples published in Su et al (2012) [2] with Replication Phase 2 samples above (section 8.4.1).

After meta-analysis, SNPs were selected for further investigation using an arbitrary cut off of $P<5x10^{-6}$. Firstly, the Linkage Disequilibrium (LD) data was assessed for all SNPs using the SNP Annotation and Proxy Search (SNAP) database (www.broadinstitute.org/mpg/snap/). If a SNP was found to be in LD ($r^2>0.4$) with another, then the SNP with the more significant *P* value was selected for genotyping at that locus. One SNP with $P<5x10^{-6}$, rs9936833, was not genotyped further in Replication Phase 3 samples, as this was a SNP previously reported by Su et al (2012) [2] in the initial BE GWAS. The remaining independent SNPs were selected for further replication in all Replication Phase 3 samples.

8.4.3 Replication Phase 3 Power Calculations

Power calculations for the Replication Phase 3 SNPs were determined using an in house power calculation tool developed by Doug Altman. The most frequently used test in genotype-based analysis for case-control genetic association studies is the Cochran-Armitage Trend test; this was therefore used in this study.

Calculations were performed for all seven Replication Phase 3 SNPs in UKREP3 (997 cases and 974 controls) and all Phase 3 replication samples (UKREP3, Belgian, Dutch Extension and BEACON; 4697 cases and 5205 controls).

8.5 STATISTICAL METHODS

8.5.1 PLINK

PLINK [142] (http://pngu.mgh.harvard.edu/~purcell/plink/) was used to check the frequency of each SNP to determine the major/minor allele. To use PLINK, a MAP and PED file were created as detailed below. Note that the columns in the PED file must match the rows in the MAP file.

8.5.1.1 MAP File

The MAP file contains 4 columns, with one SNP per line:

(i) Chromosome number

(ii) rs number

(iii) Genetic distance (morgans)

(iv) Base-pair position

An example of three SNPs can be found in Figure 8.6A.

8.5.1.2 PED File

A PED file is a space or tab delimited file. The first six columns are always as follows:

(i) Family ID

(ii) Individual ID

(iii) Paternal ID

(iv) Maternal ID

(v) Sex (1=male; 2=female; other=unknown)

A:

```
15    rs189247    0    95387634
16    rs2043633   0     5759275
12    rs2701108   0    113158644
```

B:

```
ChP102734  ChP102734  0  0  1  1  CT  TT  GG
ChP103800  ChP103800  0  0  2  1  CT  TT  AA
ChP104244  ChP104244  0  0  1  1  CC  GT  AG
ChP101440  ChP101440  0  0  1  1  CT  GT  GG
```

Figure 8.6: Files required for PLINK. MAP file example (A) for three SNPs. Columns as follows: chromosome number, rs number, Genetic distance (morgans), Base-pair position (bp units). PED file example (B) for four cases in three SNPs. Columns as follows: Family ID, Individual ID, Paternal ID (0=missing), Maternal ID (0=missing), Sex (1=male, 2=female), Phenotype (1=case, 0=control), Genotype for rs189247, Genotype for rs2043633, Genotype for rs2701108. NB. The columns in the PED file must match the rows in the MAP file.

(vi) Phenotype

Genotypes are noted in column 7 onwards. An example of a PED file can be found in Figure 8.6B. In this study, the Family and Individual ID are the same and neither Paternal nor Maternal IDs were given.

8.5.1.3 PLINK command line

An example of the PLINK command line used in this study is shown below in italics.

The following parameters were used in this research:

--file <fileroot>: specify the root filename of the PED and MAP files to be used.

--freq: output the allele frequencies.

--compound-genotypes: use AA, AG, 00 coding (no spaces between alleles in the PED file).

--out: specify output root filename

Specifically: *plink --file UKREP3_snps --freq --compound-genotypes --out /Laura/freq_BO*

8.5.2 GTOOL

GTOOL was employed to convert the MAP and PED files to GENOTYPE (GEN) and SAMPLE files, needed for SNPTEST. Firstly all missing data, denoted by -9's for the MAP and PED files, were replaced with NA's before using GTOOL.

8.5.2.1 GENOTYPE File

The GEN file stores data on each specific SNP on one line (one SNP per line). The first 5 columns are as follows:

(i) SNP ID

(ii) rs number

(iii) Base-pair position of the SNP

(iv) Allele A

(v) Allele B

The SNP ID was used for chromosome number in this study.

An example of the GEN file for one SNP in two cases can be seen in Figure 8.7A. The three numbers in columns 6, 7 and 8 show the probabilities of the three genotypes (AA, AB and BB) for the first case. The next set of three numbers show the genotype probabilities for the second case. This repeats for all cases in the GEN file.

8.5.2.2 SAMPLE File

The first three columns on the SAMPLE file are as follows:

(i) ID_1

(ii) ID_2

(iii) Missing

Each case has two ID's (columns 1 and 2) and missing data proportion (column 3). Additional entries used in this study were sex (however, this data was not analysed) and case. An example of the SAMPLE file can be found in Figure 8.7B.

A:

16	rs2043633	5759275	G	T	0	1	0	0	1	0

B:

0	0	0	D	B
ChP102227	ChP102227	0.076923	NA	1
ChP102572	ChP102572	0.384615	NA	1

Figure 8.7: Files created using GTOOL. GEN file example (A) for one SNP in two cases. The first five columns are: Chromosome number, RS number, Base-pair position of the SNP, Allele A, Allele B. The first set of three numbers show the probabilities of the three genotypes (AA, AB and BB) for the first case, the second set of three numbers show the probabilities for the second case. SAMPLE file example (B). Columns are as follows: ID_1, ID_2, missing, sex, case. Each individual can have two IDs (only one was used in this study). Sex was not included in the analyses, hence NA. The first line defines the type of variables used in each column; 0=no variable, D= Discrete, B= Binary Phenotype (eg. 0 = Controls, 1 = Cases). Both the GEN and SAMPLE file are tab-delimited files.

8.5.2.3 GTOOL command line

An example of the GTOOL command line used can be seen in italics below. The following parameters were used:

-P: PED to GEN conversion mode

--ped <filename>: specify PED format genotype file

--map <filename>: specify the MAP SNP file which accompanies the --ped file

--binary_phenotype: the phenotype in the output sample file is 'B'.

--og: output genotype file. Default, append .gen to PED file name.

--os: output sample file. Default, append .sample to PED file name.

Specifically:

gtool -P --ped UKREP3 _cases.ped --map UKREP3_ cases.map --binary_phenotype --og UKREP3_ cases.gen --os UKREP3_ cases.sample

8.5.3 Association Analysis using SNPTEST

Case-control analysis was performed using frequentist tests under a missing data logistic regression model using SNPTEST (v2.4.1).

8.5.3.1 Input Files

The input files for SNPTEST are the GEN and SAMPLE files generated using GTOOL (Section 8.5.2).

8.5.3.2 Output Files

The output file contains the following information: rs number; chromosome number; base-pair position; allele A; allele B; Effect Allele (EA); the number of cases with AA, AB or BB genotype; number of cases with missing genotype; total case number; number of controls with genotypes AA, AB or BB; controls with missing genotype; total

control number; Minor Allele Frequency (MAF) of all samples; cases MAF; controls MAF; missing data proportion; case HWE; control HWE; heterozygous OR; heterozygous OR lower; heterozygous OR upper; homozygous OR; homozygous OR lower; homozygous OR upper; OR of all samples; OR lower of all samples; OR upper of all samples; P-value; BETA; SE (Standard Error).

8.5.3.3 SNPTEST command line

An example of the command line is shown below in italics. The parameters used were as follows:

-hwe: calculate Hardy-Weinberg statistics for each cohort and combined data.

-data <a> <b>: specify data files for analysis in GEN and SAMPLE pairs (for cases and controls).

-o <a>: specify name of output file.

-frequentist <a>: specify which Frequentist tests to fit.

-method <a>: specify method used to fit model.

-pheno <a>: specify name of phenotype to use.

Specifically:

snptest2 -hwe -data UKREP3.v3.gen UKREP3.v3.sample glacier_controls.v2.gen glacier_controls.v2.sample -o /Laura/UKREP3 -frequentist 1 -method score -pheno case

The '-frequentist 1' option was used to denote an additive model. The '-method' option controls the way genotype uncertainty is taken into account when carrying out association tests. In this case, the 'score' option uses a missing data likelihood score test. The '-pheno' option specifies which phenotype to test. In this study 'case' was used (to match the SAMPLE file).

Note that SNPTEST codes allele B as the EA, to which the BETA's and SE's are based on. This needs to be recalled when analyzing SNP effects.

8.5.4 Meta-analysis using GWAMA

Meta-analysis is used to combine results from various studies and/or sample sets, increasing the statistical power to identify significant variations within a given population. Combined analysis of the data outlined in this replication study relied upon sample sets used in the BE GWAS paper (consisting of the Discovery Phase, UKREP1 and Dutch Replication) [2].

GWAMA (v2.1) is a software tool for meta-analysis of whole genome association data. It was used in this study to implement fixed inverse variance-based methods for meta-analysis [143].

8.5.4.1 Input File

The input file for GWAMA is required to have at least the following six columns for quantitative trait analysis:

(i) rs number

(ii) EA

(iii) Non Effect Allele (NEA)

(iv) P-value

(v) BETA

(vi) SE

8.5.4.2 Output File

GWAMA provides three output files. The first output file (gwama.out) contains the results and consists of the following columns: rs number; reference allele; other allele; Effect Allele Frequency (EAF); beta; SE; beta_95L (Lower 95% Confidence Interval (CI) for BETA); beta_95U (Upper 95% CI for BETA); z score; p-value; -log10 p-value; q statistic (Cochran's heterogeneity statistic); q p-value (Cochran's heterogeneity

statistic's p-value); i2 (Heterogeneity index i2); number of studies; number of samples; summary of effects direction.

The second output file (gwama.log.out) contains the log information about the current GWAMA run. Each error and warning has unique error code. More information for them can be found in the gwama.err.out file.

The gwama.err.out file is the third file produced by GWAMA and contains all errors and warnings generated during the GWAMA run.

8.5.4.3 GWAMA command line

An example of the GWAMA command line is shown below in italics. The parameters used were as follows:

--filelist <filename>: specify studies' result files. For ease, a text file containing the path directories to each input file (one file per cohort) was created for use here.

--quantitative or -qt: select quantitative trait version (BETA and SE columns).

--output <fileroot>: specify file root for output of analysis.

Specifically:

gwama --filelist Discovery_RepPhase1_RepPhase2_RepPhase3.txt -qt --output /Laura/Meta_all

8.6 IMPUTATION

Imputation describes the process of predicting genotypes that have not been directly typed in a sample of individuals. It provides a high-resolution view of a region within the genome and increases the chance that a causal SNP can be directly identified. Imputed SNPs that show greater statistical significance compared to genotyped SNPs can be better candidates for replication studies [144].

In this study, Dr Claire Palles used IMPUTE2 [145] (https://mathgen.stats.ox.ac.uk/impute/impute_v2.html) to impute 1Mb surrounding the

lead genotyped SNPs on chromosome (chr) 2p24.1 and chr12q24.21 in the Discovery data. The HapMap3 release 2 and the 2009 release of the 1000 genomes project were used as reference panels and recommended software parameters were applied.

Hit plots of the two regions were created by Dr Claire Palles using Locus Zoom, LD information used was from the 2009 release of the 1000 genomes project, CEU samples.

8.7 *IN SILICO* FUNCTIONAL ANALYSES

In silico functional analysis allows scientists to predict the effect of disease-associated SNPs. These analyses are particularly useful for intergenic and intronic SNPs, where there is no obvious functional effect.

The two new significant SNPs (rs3072 and rs2701108) were analysed using the following browsers: HaploReg V2 [146], Regulome DB [147], SNP and Copy number ANnotation (SCAN) [148], and ANNOVAR [149].

8.7.1 HaploReg

HaploReg [146] (http://www.broadinstitute.org/mammals/haploreg/haploreg.php) is an Encyclopedia of DNA Elements project (ENCODE)-funded tool that can be used for investigating possible regulatory SNPs at disease-associated loci. It was designed for scientists to aid the development of hypotheses on non-coding variants. Information on various SNPs such as chromatin state, sequence conservation, and their effect on regulatory motifs can be predicted based on LD information from the 1000 Genomes Project. This information could, in theory, provide helpful information on clinical phenotypes [146]. The parameters were set up as follows: r^2 threshold of >0.4, 1000G Phase 1 LD calculation for European population, ENCODE was used for the epigenome source, Conservation algorithm included both Genomic Evolutionary Rate Profiling (GERP) and SiPhy-omega and SNP position was shown relative to RefSeq genes.

8.7.2 RegulomeDB

RegulomeDB [147] (http://www.regulomedb.org/) is an ENCODE-funded database that annotates SNPs with known and predicted regulatory elements within introns. The database covers DNAase hypersensitivity, transcription factor binding sites, and promoters. The data present in the database comes from the Gene Expression Omnibus (GEO), ENCODE project, and published literature.

The list of SNPs in LD with each of the new BE-associated SNPs with r^2>0.4, produced by HaploReg, was entered on the RegulomeDB website. The results were then downloaded.

8.7.3 SCAN

SCAN [148] (http://www.scandb.org/newinterface/about.html) produces physical (location, flanking genes) and functional expression Quantitative Trait Loci (eQTL) annotation of SNPs. Information on physical, functional, and LD annotation on the database comes from public resources (HapMap (release 23a) and National Center for Biotechnology Information Single Nucleotide Polymorphism Database (NCBI dbSNP).

Each SNP list (the BE-associated SNP and those in LD, with r^2>0.4) was entered on the SCAN website. All default options were selected for analysis (include SNP info, include host gene and SNP function, include left- and right- flanking genes, include genes that SNP predicts expression for with *P* value less than 0.0001).

8.7.4 ANNOVAR

ANNOVAR [149] (http://www.openbioinformatics.org/annovar/) is a software tool which uses up to date information to functionally annotate genetic variants in a variety of genomes. This tool was used to perform gene-based annotations, region-based annotations and filter-based annotations. The annotations selected for the two genome-wide significant SNPs were refGene (gene-based), GERP (filter-based), PhastCons (region-based) and SiPhy (filter-based). RefGene was used to provide the location of each SNP and it's distance from nearby Refseq genes. The GERP

annotation was used to provide a GERP score of conservation, which ranges from -12.3 to 6.17, with 6.17 being the most conserved. The PhastCons annotation was used to assess conservation of genomic regions. PhastCons scores range from 0-1000, with 1000 being the most conserved. Finally, the SiPhy annotation was employed to assess conservation across 29 mammalian genomes; the larger the score, the more conserved the site.

8.7.4.1 Input File

An input file for each SNP list (the BE-associated SNP and those in LD, with r^2>0.4) was produced for ANNOVAR. Each tab-delimited input file (.avinput) required the following columns:

(i) Chromosome

(ii) Start position

(iii) End position

(iv) Reference nucleotide

(v) Observed nucleotide

8.7.4.2 Files Downloaded for Analysis

In order to annotate each SNP list with the gene-, region- and filter-based annotations listed above, four files from the ANNOVAR website were downloaded as follows, note that the files used were in relation to hg19:

(i) hg19_refGene.txt: FASTA sequences for all annotated transcripts in RefSeq Gene

(ii) hg19_gerp++gt2.txt: whole-genome GERP++ scores greater than 2 (RS score threshold of 2 provides high sensitivity while still strongly enriching for truly constrained sites).

(iii) hg19_phastConsElements46way: PhastCons Scores

(iv) hg19_ ljb23_siphy: whole-exome SiPhy scores (version 2.3).

An example of the command line used to download the GERP++ file, using a perl script, from ANNOVAR can be seen below in italics. '-downdb' denotes download database, '-buildver <version>' specify which build version to download (hg19 in this case), '-webfrom <source> <file> <output location>' specify the download source, file to download and output location of that file.

perl annotate_variation.pl -downdb -buildver hg19 -webfrom annovar gerp++gt2 humandb/

8.7.4.3 Analysis

Due to the presence of more than one annotation, the table_annovar perl script was employed for analysis. This allows a one-line command to produce multiple annotations per input file and output a comma-separated value file. The .csv file can be viewed in excel, where each annotation is shown in a separate column. The parameters used were as follows:

table_annovar.pl <input file>: defines which perl script to use and location and name of input file

humandb/: defines the location of the protocol files (refGene, PhastCons, GERP and SiPhy)

-buildver <version>: specify the build version

-out <filename>: specify the output file name

-remove: denotes removal of all temporary files produced before configuring the .csv output file

-protocol <a,b,c,d>: specify the annotation protocol files to use

-operation<a,b,c>: specify the operation of the protocols used (g=gene-based, r=region-based, f=filter-based)

-nastring <a>: specify what to insert in the output file if the annotation is empty

-csvout: denotes the csv output file.

Specifically:

perl table_annovar.pl example/rs2701108.avinput humandb/ -buildver hg19 -out rs2701108 -remove -protocol refGene,phastConsElements46way,gerp++gt2,ljb23_siphy -operation g,r,f,f -nastring NA –csvout

8.7.4.4 Output File

The output file consisted of the following columns:

(i) Chromosome

(ii) Start postition

(iii) End postition

(iv) Reference Allele

(v) Alternative Allele

(vi) Function according to refGene

(vii) Nearby Genes according to refGene

(viii) Gene Detail (distance of SNP from gene) from refGene

(ix) Exonic Function according to refGene

(x) Amino acid change according to refGene

(xi) PhastCons score according to phastConsElements46way

(xii) GERP score according to gerp++gt2

(xiii) SiPhy score according to ljb23_siphy

An example of the output file in excel is shown in Table 8.4, showing the first three

Chr	Start	End	Ref	Alt	Func.refGene	Gene.refGene	GeneDetail. refGene	ExonicFunc. refGene	AAChange. refGene	phastConsEle ments46way	gerp++gt2	ljb23_siphy
12	114660658	114660658	T	C	intergenic	RBM19,TBX5	dist=256482; dist=131077	NA	NA	Score=556; Name= lod=243	NA	NA
12	114661066	114661066	C	G	intergenic	RBM19,TBX5	dist=256890; dist=130669	NA	NA	NA	2.1	NA
12	114661789	114661789	A	T	intergenic	RBM19,TBX5	dist=257613; dist=129946	NA	NA	NA	NA	NA

Table 8.4: Example of the ANNOVAR functional analysis, showing the output file for the first three SNPs of the rs2701108 analysis, in Microsoft Excel. ANNOVAR was used to annotate SNPs in LD (r^2>0.4) with the new BE-associated SNPs. The ANNOVAR table_annovar perl script was used with the following protocols, downloaded from the ANNOVAR website, for SNP annotation: hg19_refGene.txt, hg19_gerp++gt2.txt, hg19_phastConsElements46way.txt and ljb2_siphy.txt.

SNPs of the rs2701108 input file analysis.

8.8 ASSOCIATION TESTING OF THE FOUR BE SNPS IN EAC CASES

BE predisposes patients to EAC. It was therefore decided to explore the link between the two diseases.

8.8.1 Adenocarcinoma Case Selection

Case phenotypes reported as Siewert type 1 adenocarcinomas without any evidence of BE (therefore EAC-only) were used in this section of the study. Siewert type 2 or 3 adenocarcinomas were not used due to the close proximity to the stomach, and hence possible misdiagnosis of stomach cancer.

8.8.2 Sample Set

An independent set of 305 UK and 176 Dutch cases (481 in total) were used. These samples were collected as part of ChOPIN.

The controls used in this stage were the same as those used in Replication Phase 1 of the BE analysis (UK N=6,819 from PoBI and 58C; Dutch N=1,780 from University Medical Centre, Groningen).

8.8.3 Genotyping

All gDNA had been extracted previously under the BE GWAS, but had not been analysed. The genotypes for the selected SNPs, detailed below, were extracted from the Immunochip.

8.8.4 SNP Selection

Four SNPs associated with BE (the two previously identified SNPs: rs9257809 and rs9936833; and the two newly identified SNPs: rs3072 and rs2701108) were selected for analysis.

8.8.5 Analysis

Association testing was performed for a case-control analysis using the EAC-only cases and Replication Phase 1 controls. In addition to this, a case-only meta-analysis,

using BE cases from Replication Phase 1 and the EAC-only cases outlined here, was performed consisting of; UK 2,957 BE vs 305 EAC-only and Dutch 473 BE vs 176 EAC-only.

8.9 REPLICATION OF BE/EAC SNPS REPORTED BY LEVINE ET AL (2013) [3]

A combined BE/EAC GWAS paper was published by Levine et al (2013) [3] (comprising 2,363 EAC cases, 3,116 BE cases and 10,060 controls) whilst completing this replication study [3]. The paper analysed 922,031 SNPs in the Discovery phase and 87 of 94 SNPs with $P<1\times10^{-4}$ in the replication phase. Levine et al (2013) [3] identified four SNPs (three loci) associated with BE/EAC risk: rs2687201, rs11789015, rs10419226, and rs10423674.

8.9.1 Controls

The controls for each sample set was the same as in the BE analysis apart from the Discovery phase controls. The WTCCC2 controls used in our original Discovery phase overlapped with those in Levine et al (2013) [3], hence new controls from colorectal cancer GWAS studies CORGI and Colon Cancer Family Registry (CFR) were used [127, 150]. The details of CORGI are available in section 8.1.2. The CRF cohort is an international consortium of six institutes in North America and Australia, which provides support to studies on the etiology, prevention, and clinical management of colorectal cancer. The cohort comprises data and specimens from over 40,000 participants from 14,000 families that have been recruited from 1998 to 2011 [150].

8.9.2 Replication Analysis of the four BE/EAC-associated SNPs

8.9.2.1 Cases

The cases included in this replication study varied depending on the SNP analysed.

Three of the four SNPs (rs2687201, rs11789015 and rs10419226) were not directly genotyped in our samples. Therefore, they were imputed (see section 8.9.4) in only our

Discovery phase, resulting in SNP association testing in 3 sample sets overall: the Levine Discovery, Levine Replication and our Discovery.

However, rs10423674 was directly genotyped in our Discovery, UKREP1 and Dutch replication samples, and so was analysed in 5 sample sets overall: the Levine Discovery, Levine Replication, Our Discovery, UKREP1 and Dutch Replication.

8.9.3 Replication Analysis of the remaining 83 SNPs

8.9.3.1 SNP selection

From the 87 BE/EAC SNPs from Levine et al (2013) [3], 83 were yet to be analysed in our data. Only 10 of these 83 were not genotyped or reliably imputed (see section 8.9.4) with an info score >0.95. One of the 10 SNPs, rs11771429, had $P<10^{-5}$, consequently, this SNP was genotyped using the KASP genotyping technology (section 8.3.2.4) in our Discovery phase samples. Therefore, 74 SNPs were included in the meta-analysis of Levine's Discovery, Levine's Replication and our Discovery. Four SNPs were selected for further replication based on *P* value and LD information: rs1497205, rs254348, rs3784262 and rs4523255.

8.9.3.2 Cases

The cases included in this stage varied depending on the SNP.

All four SNPs (rs1497205, rs254348, rs4523255 and rs3784262) were genotyped in 6 cohorts: the Levine Discovery, Levine Replication, our Discovery, UKREP1, Dutch Replication (including Extension) and UKREP2.

After meta-analysis at this stage one SNP (rs3784262) was the most significant and was therefore additionally genotyped in the Irish Replication, UKREP3 and Belgian Replication samples (total number of cohorts for rs3784262=9).

8.9.3.3 Genotyping

SNPs selected for genotyping in Replication Phases 1, 2 and/or 3 were genotyped using the KASP genotyping technology (section 8.3.2.4) [141].

8.9.4 Imputation

Imputation was completed by Dr Claire Palles. To compensate for inter-study differences in array content, SNP genotypes were phased using SHAPEIT (to determine the haplotypes; which allele belongs to which copy of a specific chromosome or which alleles appear together on the same chromosome) and imputed (to estimate genotypes of individuals with missing data, using known haplotypes) using IMPUTE2 using recommended software options. The September 2013 release of the 1000 genomes project was used as a reference panel. SNPs with IMPUTE2 info scores of <0.95 or showing departures from Hardy-Weinberg equilibrium ($P<10^{-6}$) were excluded.

8.9.5 Principle Component Analysis

Principle Component Analysis (PCA) of our amended Discovery set (1,852 AspECT cases and 1,898 controls from CFR1 and CORGI), completed by Dr Claire Palles, was used to remove outlying samples, leaving 1,741 cases and 1,642 controls for analysis. λ_{GC} was 1.077 prior to the inclusion of the first principle component (PC) and 1.068 after adjustment, suggesting that population structure was not a major confounder in our discovery phase.

8.9.6 Association Testing and Meta-analysis

Association testing and meta-analysis for the Levine SNPs was completed in the same way detailed in section 8.5, using SNPTEST and GWAMA [143].

9. RESULTS

As outlined in methods, Discovery and Replication Phases 1 and 2 sample sets within this project have been used in the initial BE GWAS paper by Su et al (2012) [2]. Discovery, UKREP1, Dutch Replication, UKREP2 and the Irish Replication have been used previously to identify two BE SNPs (rs9257809; $P=4.09\times10^{-9}$, OR=1.21 and rs9936833; $P=2.74\times10^{-10}$, OR=1.14). However, the Replication Phase 2 sample sets were only used to validate two SNPs in the previous GWAS. In this replication study, the SNP selection criteria for Replication Phase 2 samples were relaxed to $P<10^{-4}$ rather than $P<10^{-8}$, resulting in analysis of 83 SNPs. An overview of the sample sets and number of SNPs analysed is shown in Figure 9.1.

9.1 IDENTIFICATION OF TWO NOVEL BE SNPS

9.1.1 SNPs Prioritised for Replication in Phase 2 Samples

Three custom Sequenom iPLEXes were designed by Dr Claire Palles at the WTCHG. A total of 83 SNPs were prioritised for further genotyping based on the following criteria:

(i) $P<10^{-4}$ in combined Discovery and Replication Phase 1 analysis in Su et al (2012) [2] (N=63)

(ii) $P<10^{-4}$ in Discovery Phase, but not included in Replication Phase 1 in Su et al (2012) [2] (N=12)

(iii) $P<10^{-4}$ in a sex-stratified analysis of the Discovery phase (N=5)

(iv) Candidate polymorphisms previously reported as associated with BE and not well tagged by the Discovery Phase or Replication Phase 1 (N=3)

Assays were successfully designed for 65 of the 83 SNPs (details can be found in Section 8.4.1). These SNPs were genotyped in the Replication Phase 2 samples; UKREP2 set (1,765 cases and 1,586 controls) and the Irish set (245 cases and 473 controls). A meta-analysis of the Discovery and Replication Phases 1 and 2 was completed by Dr Claire Palles (results produced by Dr Claire Palles can be found in the

appendix, Table 11.1). A selection of these SNPs (N=12) were selected for replication in Phase 3 samples.

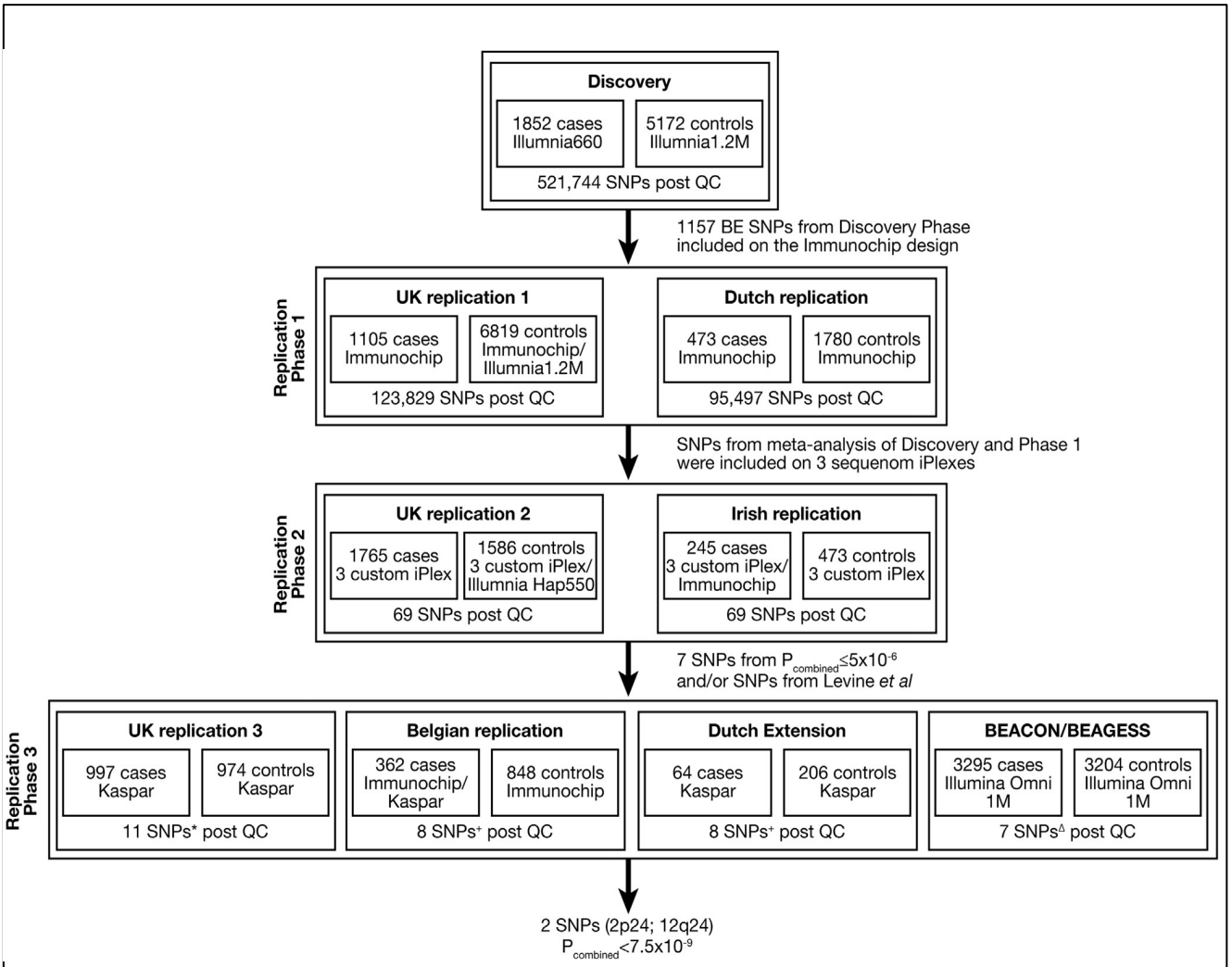


Figure 9.1: Outline of the phases and the SNPs analyzed [1]. Discovery and Replication Phase 1 were used in the previous GWAS by Su et al [2]. Two SNPs described in Su et al [2] had previously been genotyped in Replication Phase 2 and BEACON/BEAGESS samples. Replication Phase 3 samples are new to this replication study. Dutch Replication (Phase 1 Replication) and the Dutch extension (Phase 3 Replication) is one cohort in our analyses for the SNPs taken through to Replication Phase 3. *11 SNPs: Our SNPs; rs3072, rs6751791, rs2731672, rs2701108, rs189247, rs2043633, and rs12985909 and Levine et al [3] SNPs; rs1497205, rs254348, rs3784262, and rs4523255. $^{+}$8 SNPs: Our SNPs rs3072, rs6751791, rs2731672, rs2701108, rs189247, rs2043633, and Levine et al SNP: rs3784262. $^{\Delta}$7 SNPs: Our SNPs rs3072, rs6751791, rs2731672, rs2701108, rs189247, rs2043633.

9.1.2 SNPs Prioritised for Replication in Phase 3 Samples

SNPs selected for genotyping in Replication Phase 3 samples were based on a meta-analysis between Discovery set and Replication Phase 1 samples published in Su et al (2012) [2] with Replication Phase 2 samples. Following meta-analysis, 12 SNPs met the arbitrary cut off of $P<5x10^{-6}$, shown in Table 9.1. Five SNPs were excluded from further replication. rs7255 had a call rate <95% on the Sequenom iPLEX, and was also in LD ($r^2=0.51$) with rs3072 (which had a more significant *P* value at this stage), so was excluded. rs9936833 was excluded from replication as it is a previously reported BE susceptibility SNP [2]. rs12993283 was excluded as it was in complete LD with rs6751791 ($r^2=1.00$).

SNP	CHR	Position	Meta OR	Excluded	Reason for Exclusion
rs3072	2	20741887	1.15	No	-
rs2043633	16	5759275	1.15	No	-
rs7255	2	20742301	1.15	Yes	<95% call rate and in LD with rs3072 ($r^2=0.51$)
rs9936833	16	84960619	1.15	Yes	Previously reported BE susceptibility SNP
rs189247	15	95387634	1.14	No	-
rs2701108	12	1.13E+08	1.14	No	-
rs6751791	2	35435501	1.14	No	-
rs12993283	2	35445909	1.12	Yes	In LD with rs6751791 ($r^2=1.00$)
rs9941024	16	5734313	1.12	Yes	In LD with rs2043633 ($r^2=0.75$)
rs2731672	5	1.77E+08	1.14	No	-
rs12985909	19	18300383	1.11	No	-
rs11866983	16	5743925	1.12	Yes	In LD with rs2043633 ($r^2=0.42$)

Table 9.1: SNPs selected for Phase 3 Replication, based on meta-analysis between Discovery and Replication Phases 1 and 2. Twelve SNPs, ranked by *P*-value, with $P<5x10^{-6}$ after meta-analysis between Discovery and Replication Phases 1 and 2. Five SNPs were excluded from Replication Phase 3. Chr, Chromosome; LD, Linkage Disequilibrium; OR, Odds Ratio.

rs9941024 and rs11866983 were in LD with rs2043633 (r^2=0.75 and 0.42 respectively), hence only rs2043633 was carried forward to Replication Phase 3 (as it had the strongest association after Discovery and Replication Phases 1 and 2 meta analysis).

The seven remaining SNPs (rs3072, rs2043633, rs189247, rs2701108, rs6751791, rs2731672, rs12985909) were therefore genotyped in all Replication Phase 3 samples; UKREP3 (997 cases and 974 controls), Dutch Extension (64 cases and 206 controls), Belgian (362 cases and 848 controls) and BEACON (3295 cases and 3204 controls) sets.

9.1.3 Power Calculations for UKREP3 and Replication Phase 3

Power calculations are an important tool when determining the probability that a statistical significance test will reject the null hypothesis. Within this study, power calculations were determined using the Cochran-Armitage Trend test, as this is the most frequently used in genotype-based tests for case-control genetic association studies. Results from these calculations in UKREP3 (997 cases and 974 controls) and all Phase 3 replication samples (4697 cases and 5205 controls) can be seen in Table 9.2. From Table 9.2, it is clear that there is not significant power (>0.80) to determine a

SNP	CHR	Position	Discovery MAF (cases/controls)	Effect Size*		Case:control ratio+	Cochran-Armitage Trend Test
rs3072	2	20878406	0.41/0.36	1.16	**UKREP3**	1.02:1	0.62
					Rep3	0.90:1	1.00
rs6751791	2	35581997	0.51/0.48	1.13	**UKREP3**	1.02:1	0.48
					Rep3	0.90:1	0.99
rs2731672	5	176842474	0.27/0.24	1.14	**UKREP3**	1.02:1	0.43
					Rep3	0.90:1	0.98
rs2701108	12	114674261	0.38/0.41	1.14	**UKREP3**	1.02:1	0.53
					Rep3	0.90:1	1.00
rs189247	15	97586630	0.41/0.37	1.14	**UKREP3**	1.02:1	0.52
					Rep3	0.90:1	0.99
rs2043633	16	5819274	0.37/0.41	1.15	**UKREP3**	1.02:1	0.58
					Rep3	0.90:1	1.00
rs12985909	19	18439383	0.48/0.45	1.12	**UKREP3**	1.02:1	0.43
					Rep3	0.90:1	0.98

Table 9.2: Power calculations, for the seven SNPs taken through to Replication Phase 3, in UKREP3 samples and all Replication Phase 3 samples. Position is based on build 37. *Effect Size was determined by meta-analysis of Discovery, Replication 1 and Replication 2. + The case:control ratio stated was determined in the Discovery phase. UKREP3 consisted of 997 cases and 974 controls. Replication Phase 3 (Rep3) consisted of 4697 cases and 5205 controls. MAF, Minor Allele Frequency.

true variation when using the UKREP3 sample set. The SNP with greatest power is rs3072 (power = 0.62). However, when all Replication Phase 3 samples are used, there is significant power to determine a true variation for all seven SNPs. This highlights the importance of sample size.

9.1.4 Meta-analysis of Discovery and the three Replication Phases

SNP association analysis of the seven SNPs in only the UKREP3 sample set (997 cases and 974 controls) can be seen in Table 9.3. rs3072 is significant in the single UKREP3 set with *P*=0.018, OR=1.18, 95%CI=1.04-1.34, all other SNPs show no significance in this sample set (*P*<0.05). This result is not unexpected, since the power calculation (section 9.1.3) showed that rs3072 was the most likely (out of the seven SNPs tested) to show significance in the UKREP3 sample set, as it had the greatest power.

9.1.4.1 Corrections for Multiple Testing

A *P* value threshold of 0.05 has been accepted as a reasonable level for declaring significance. When this threshold is enforced, the chance of making a Type I error

SNP	CHR	Position	EA	OR (95%CI)	*P*
rs3072	2	20878406	G	1.18 (1.04-1.34)	$1.18x10^{-2}$
rs6751791	2	35581997	G	1.02 (0.86-1.20)	$8.04x10^{-1}$
rs2731672	5	176842474	A	1.02 (0.85-1.24)	$8.13x10^{-1}$
rs2701108	12	114674261	A	1.03 (0.91-1.16)	$6.69x10^{-1}$
rs189247	15	97586630	A	1.07 (0.94-1.22)	$2.80x10^{-1}$
rs2043633	16	5819274	C	1.03 (0.90-1.17)	$6.83x10^{-1}$
rs12985909	19	18439383	G	1.03 (0.91-1.16)	$6.47x10^{-1}$

Table 9.3: SNP association analysis of the seven SNPs in the UK Replication 3 sample set consisting of 997 Barrett's Oesophagus cases and 974 female controls. Results are presented with respect to the effect allele. Chr, Chromosome; EA, Effect Allele; OR, Odds Ratio; 95%CI, 95% Confidence Intervals; *P*, *P value.*

(falsely claiming significance) is 5% ($\alpha=0.05$). This 5% error rate applies to each statistical test run. Therefore, the more analyses run, the higher the chance of Type I error.

One way to correct for multiple testing in GWAS is the Bonferroni correction. The Bonferroni correction adjusts alpha to 0.05/n, where n is the number of statistical tests performed. Most GWAS use genotyping platforms which test for 1,000,000 SNPs; hence the statistical significance threshold of a single SNP association in a GWAS is set at 5×10^{-8} [151-154].

9.1.4.2 Meta-analysis

Meta-analysis of the Discovery and the three Replication phases, seen in Table 9.4, identified two intergenic SNPs associated with BE risk, from a total sample size of 10,158 BE cases and 21,062 controls. rs3072 on chr2p24: $P_{meta}=1.8\times10^{-11}$; OR=1.14; 95%CI=1.09-1.18 and rs2701108 on chr12q24: 7.5×10^{-9}; OR=1.11; 95%CI=1.08-1.16.

SNP	CHR	Position	EA	OR (95%CI)	*P*
rs3072	2	20878406	G	1.14 (1.09-1.18)	1.75×10^{-11}
rs6751791	2	35581997	A	1.08 (1.04-1.12)	7.65×10^{-5}
rs2731672	5	176842474	A	1.07 (1.03-1.12)	$1.66\text{x}10^{-3}$
rs2701108	12	114674261	A	1.11 (1.08-1.16)	7.48×10^{-9}
rs189247	15	97586630	A	1.10 (1.06-1.14)	3.55×10^{-7}
rs2043633	16	5819274	A	1.09 (1.05-1.14)	2.25×10^{-6}
rs12985909	19	18439383	G	1.10 (1.06-1.14)	3.28×10^{-7}

Table 9.4: Final meta-analysis of all sample sets for the seven SNPs taken through to Replication Phase 3. Results are presented with respect to the effect allele. Position is based on build 37. Results based on meta-analysis of all sample sets comprising UK Discovery, UK Replication 1, Dutch Replication and Extension, Irish Replication, UK Replication 2, UK Replication 3, Belgian Replication and the BEACON Replication. Chr, Chromosome; EA, Effect Allele; OR, Odds Ratio; 95%CI, 95% Confidence Intervals; *P*, *P value* after meta-analysis.

9.1.5 Restricting cases to Intestinal Metaplasia for the two novel SNPs

The standard UK criteria, in accordance with the British Society of Gastroenterology, for the diagnosis of BE was used throughout this study [53]. However, the American College of Gastroenterology criteria are used in other countries, which requires the presence of IM for diagnosing BE. Therefore, in order to investigate the effect of including non-IM cases, the meta-analysis was restricted to IM-only individuals (N=8,521). The results show that the associations remained at or near genome-wide significance ($P<5\times10^{-8}$; rs3072: $P=1.3\times10^{-9}$, OR=1.13, 95%CI=1.09-1.17; rs2701108: $P=6.2\times10^{-8}$, OR=1.11, 95%CI=1.08-1.16) when compared to the meta-analysis comprising all IM positive and negative cases (rs3072: $P_{meta}=1.8\times10^{-11}$, OR=1.14, 95%CI=1.09-1.18; rs2701108: 7.5×10^{-9}; OR=1.11; 95%CI=1.08-1.16).

9.1.6 Imputation of SNPs within 1Mb of the novel SNPs

To assess the disease-association strength of the genotyped SNPs, the Discovery phase data was imputed for all SNPs within 1Mb each side of the chr2p24 and chr12q24 intergenic hits (Figure 9.2). Imputation of our data, completed by Dr Claire Palles, showed that at chr2p24, rs3072 remained the most significant SNP. However, at chr12q24, rs1920562 ($r^2=0.6$) was more significant ($P_{Discovery}=1.4\times10^{-5}$, OR=0.84) compared to the lead genotyped SNP ($P_{Discovery}=1.4\times10^{-3}$, OR=0.88).

9.1.7 Identification of nearby Genes

Using the Ensembl Genome Browser (http://www.ensembl.org), it was noted that rs3072 lies 7.5kb downstream of *GDF7* (Growth Differentiation Factor 7, also known as *BMP12; Bone Morphogenetic Protein 12*) and 6.5kb downstream of *C2orf43* (chromosome 2 open reading frame 43). rs2701108 lies 117kb downstream of *TBX5* (T-Box 5; transcription factor) and 270kb upstream of *RBM19* (RNA binding motif protein 19) (Figure 9.2). Whilst the imputed, more significant SNP at this locus, rs1920562, lies 131kb downstream of *TBX5* and 256kb upstream of *RBM19*.

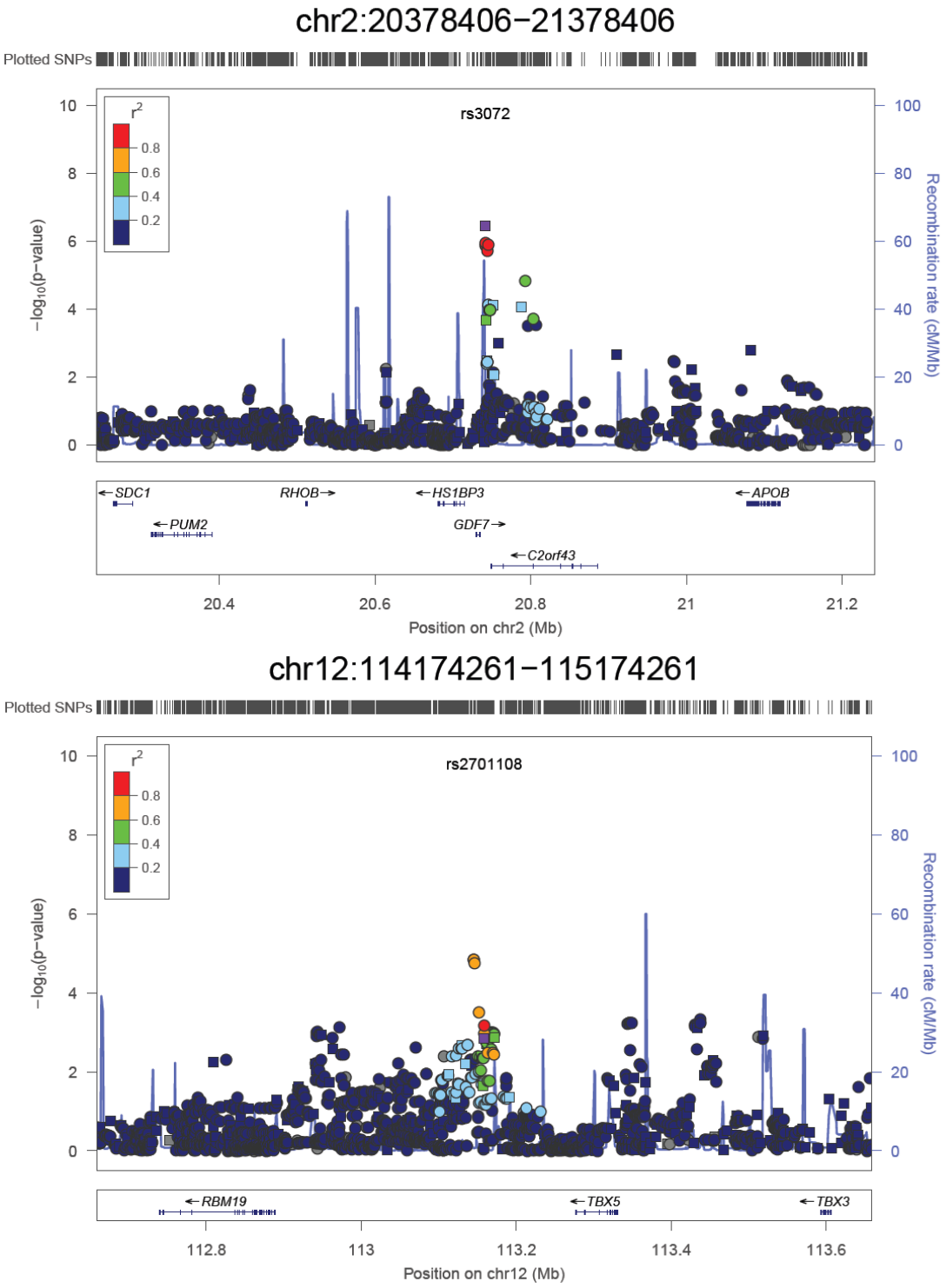


Figure 9.2: Regional plots of association (left y-axis) and recombination rates (right y-axis) for the chromosomes 2p24 and 12q24 loci following imputation [1]. rs3072 remains most significant in the chr2 region, but rs1920562 is more significant than rs2701108 in the chr12 region. The lead genotyped SNP is marked with a purple square. Imputed SNPs are plotted as circles and genotyped SNPs as squares.

9.1.8 Assessment of Non-Synonymous SNPs within nearby Genes

Non-synonymous variants (mutations which alter the amino acid sequence of a protein) in the genes near the SNPs on chr2 and chr12 were not in strong LD (r^2<0.4) with the newly identified SNPs (rs3072 and rs2701108). This suggests that the novel SNPs may not have associations with known non-synonymous SNPs, but they could have an effect on gene expression/regulation rather than the protein sequence itself.

9.1.9 *In silico* Functional Analysis of the Two Novel BE-associated SNPs

There are a variety of databases that can be used to predict function and effect of intronic SNPs. For the two novel BE-associated SNPs (rs3072 and rs2701108), HaploReg [146], RegulomeDB [147], Annovar [149] and SCAN [148] were interrogated. In order to gain a true sense of function or effect of SNPs at these loci, variants in LD with the two novel BE SNPs (r^2>0.4) were determined using the HaploReg database. Each SNP list was used to determine predicted effects and sequence conservation of the two novel SNPs.

9.1.9.1 rs3072 Analysis

The first BE-associated SNP, rs3072, is located in a region of histone modifications that marks enhancers, such as H3K4Me1, between genes *GDF7* and *C2orf43* (as can be seen in Figure 9.3 using the UCSC Genome Browser [5]). This data was collected from the lymphoblastoid cell line (LCL), GM12878. Histone modifications, such as H3K4Me1, are associated with active regions (i.e. involved in/important for transcription) of the genome; H3K4Me1 in particular is one of the most robust epigenetic marks that appears to be essential for the regulation of different cellular processes [155]. According to HaploReg, rs3072 was listed as having the potential to alter a GATA binding motif (a family of transcription factors). Details of the motif change can be seen in Table 9.5 (Table 9.5B shows the nucleotide coding used in Table 9.5A). However, RegulomeDB found “minimal binding evidence” for the bound protein RFX3, a transcription factor important in the regulation of the expression of genes involved in ciliary assembly and motility [156].

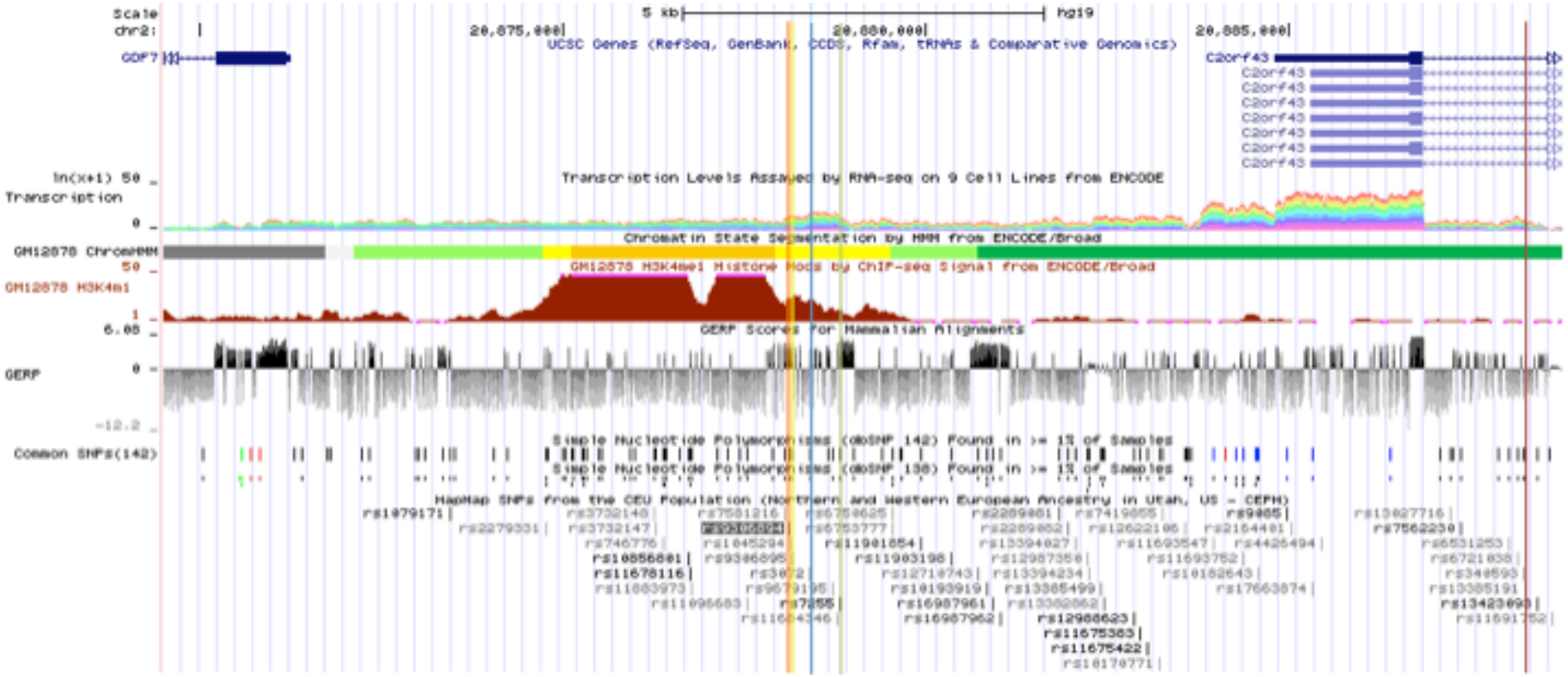


Figure 9.3: Diagrammatic representation of rs3072 *in silico* analysis using the UCSC Genome Browser. The UCSC Genome browser (hg19 build; https://genome.ucsc.edu [5]) was interrogated for rs3072 (indicated by a blue vertical line) and SNPs in LD (rs9306894, orange vertical line; rs9306895, yellow vertical line; rs7255, green vertical line; rs13385191, red vertical line). Four SNPs (rs3072, rs9306894, rs9306895 and rs7255) lie within an enhancer region predicted by the ENCODE project, marked by H3K4Me1 histone modifications. Only rs7255 mapped to a site of evolutionary conservation (GERP=2.28) within this enhancer region. One SNP, rs13385191 (red vertical line) lying within an intron of *C2ORF43*, also mapped to a conservation site with a GERP score of 2.86.

A:

rs3072

Regulatory Motif	Strand	Ref: TGCAGCTTAGAAAGCAAATTTCATCTGAT**T**CCAGTACTGTGATTTTAAGGAAACGGTAA Alt: TGCAGCTTAGAAAGCAAATTTCATCTGAT**C**CCAGTACTGTGATTTTAAGGAAACGGTAA
GATA	+	ABCTGAT**M**

rs9306894

Regulatory Motif	Strand	Ref: TTTCTTTGACGAAGAATCATAATTGAGTC**A**CTTTAGGTCTTTTAGCTGGAAGCATTTCA Alt: TTTCTTTGACGAAGAATCATAATTGAGTC**G**CTTTAGGTCTTTTAGCTGGAAGCATTTCA
AP-1	-	WMKKAGTC**A**BY
AP-1	-	TGAKTC**A**
AP-1	+	**V**TGACTHA

B:

Nucleotide code	Possible Nucleotides
R	A or G
Y	C or T
K	G or T
M	A or C
S	G or C
W	A or T
B	G or C or T
D	A or G or T
H	A or C or T
V	A or C or G

Table 9.5: Regulatory motif changes predicted by HaploReg for the BE susceptibility SNP rs3072 and one SNP in LD, rs9306894 (r^2=0.97). (A) Each SNP is denoted by **bold** font. rs3072 is predicted to alter a GATA binding motif according to HaploReg (top). rs9306894, according to both HaploReg and ReglomeDB, has the potential to alter the binding motif of Activator Protein 1 (AP-1; bottom). (B) The ambiguous nucleotide codes are specified in Table9.4B.

Three additional SNPs in LD with rs3072 (rs9306894, r^2=0.97; rs9306895, r^2=0.97; and rs7255, r^2=0.60) map to the enhancer region detected in GM12878, as can be seen in Table 11.5. Using ANNOVAR, GERP and PhastCons scores showed that rs7255 mapped to a site of high evolutionary conservation (GERP=2.28, PhastCons=501), suggesting important function for this base/region. Neither of the other two SNPs within the same enhancer region showed conservation. RegulomeDB showed that rs9306894,

although not at a conserved site, is predicted as “likely to affect protein binding and linked to expression of a gene target”. This database also showed that *C2orf43* is the most likely target of rs9306894 following eQTL studies in monocytes. Both RegulomeDB and HaploReg show that rs9306894 also has the potential to affect an AP-1 binding motif (a transcription factor important in the regulation of proliferation, differentiation and apoptosis). rs9306895 has “minimal binding evidence” according to RegulomeDB. Neither database could agree on a possible change in binding motif for this SNP. None of the three SNPs were associated with *GDF7* (the nearest protein coding gene) expression.

Another SNP in LD with rs3072, rs13385191 (r^2=0.46), achieved a GERP score of 2.86 and had “minimal binding evidence” according to RegulomeDB. SCAN showed that this SNP was located within an intron of *C2orf43.* RegulomeDB also showed that *C2orf43* is the most likely target of this SNP following eQTL studies in fibroblasts.

The Human Protein atlas shows the level of *GDF7* RNA expression at 1FPKM, but protein expression was not recorded. It also shows that *C2orf43* RNA expression is 14FPKM and the protein is expressed at a medium level in the oesophagus.

An overview of the data can be found in Figure 9.3 and in the appendix (Table 11.5).

9.1.9.2 rs2701108 Analysis

Analysis of the second BE-associated SNP, rs2701108 (located between genes *TBX5* and *RBM19*), using three of the four databases showed that it is not likely to be a functionally regulatory SNP itself. However, HaploReg did predict rs2701108 to affect the binding of the TBX5 protein (Table 9.6). When comparing the genotyped SNP to those in LD, it appears as though rs1920562 (r^2=0.62), the SNP that showed the strongest signal upon imputation of the chr12q24 loci (section 9.1.6), is a more promising candidate. rs1920562 maps to a highly conserved base (PhastCons score = 556) and a region containing enhancer marks in human embryonic stem cells (h1-ESC) and lung fibroblasts (NHLF), as seen in Figure 9.4.

rs2701108

Regulatory Motif	Strand	Ref: ATTCAGGCAGGAGAAAATGTGTACTCTCA**T**CTTTTCCAGAGTCACCTAGGAGGCAGGGC Alt: ATTCAGGCAGGAGAAAATGTGTACTCTCA**C**CTTTTCCAGAGTCACCTAGGAGGCAGGGC
TBX5	-	YTCACA**C**CTK

rs1950090

Regulatory Motif	Strand	Ref: CAGCTACAATGGGCATGTGACTCAGGGTG**A**CCAATCACAGCTCCTCCCTGCCTTAGGGA Alt: CAGCTACAATGGGCATGTGACTCAGGGTG**G**CCAATCACAGCTCCTCCCTGCCTTAGGGA
NF-Y	-	NBYR**R**CCAATSRRMR
NF-Y	+	VBBR**R**CCAATSRSVDN
NF-Y	+	DBTA**R**CCAATCARD

rs12828548

Regulatory Motif	Strand	Ref: AAGTGGAGATCCAATTCCTTCTTCCAATA**C**GTTTGTTATTTCTAAATAGCAACTCAGCT Alt: AAGTGGAGATCCAATTCCTTCTTCCAATA**G**GTTTGTTATTTCTAAATAGCAACTCAGCT
FOXJ1	+	HWD**T**GTTTGTTTA
RAD21	-	DMCACYA**G**GT
ZFP105	-	**H**NHWTKTTDWTTRHD

rs1920568

Regulatory Motif	Strand	Ref: TCTCCTAGGAACTGTAAGTTATGTCCTTC**G**TGCCTTTCTGAAAACCATCTTAAGCTGTT Alt: TCTCCTAGGAACTGTAAGTTATGTCCTTC**C**TGCCTTTCTGAAAACCATCTTAAGCTGTT
EWSR1-FLI1	-	CCTTCCTTC**C**TTCCTTCC
ETS	-	DBKNRCWTC**C**KSYBHN
STAT	-	WTC**C**TSCCT

Table 9.6: Regulatory motif changes predicted by HaploReg for the BE susceptibility SNP rs2701108 and three SNPs in LD, rs1950090 (r^2=0.42), rs12828548 (r^2=0.62) and rs1920568 (r^2=0.58). Each SNP is denoted by **bold** font. rs2701108 is predicted to affect the binding of Transcription Box 5 (TBX5) according to HaploReg. rs1950090 is predicted to alter the binding motif of Nuclear Factor Y (NF-Y) according to both RegulomeDB and HaploReg. rs12828548 is predicted to alter the binding motifs of Forkhead box protein J1 (FOXJ1), RAD21 and Zinc Finger Protein 105 (ZFP105). rs1920568 is predicted to alter the motif of EWSR1/FLI1 (Ewing sarcoma breakpoint region 1/Friend leukemia integration 1 transcription factor; a chimeric fusion oncogene), E26 transformation-specific (ETS) and Signal Transducer and Activator of Transcription (STAT).
The ambiguous nucleotide codes are specified in Table 9.4B.

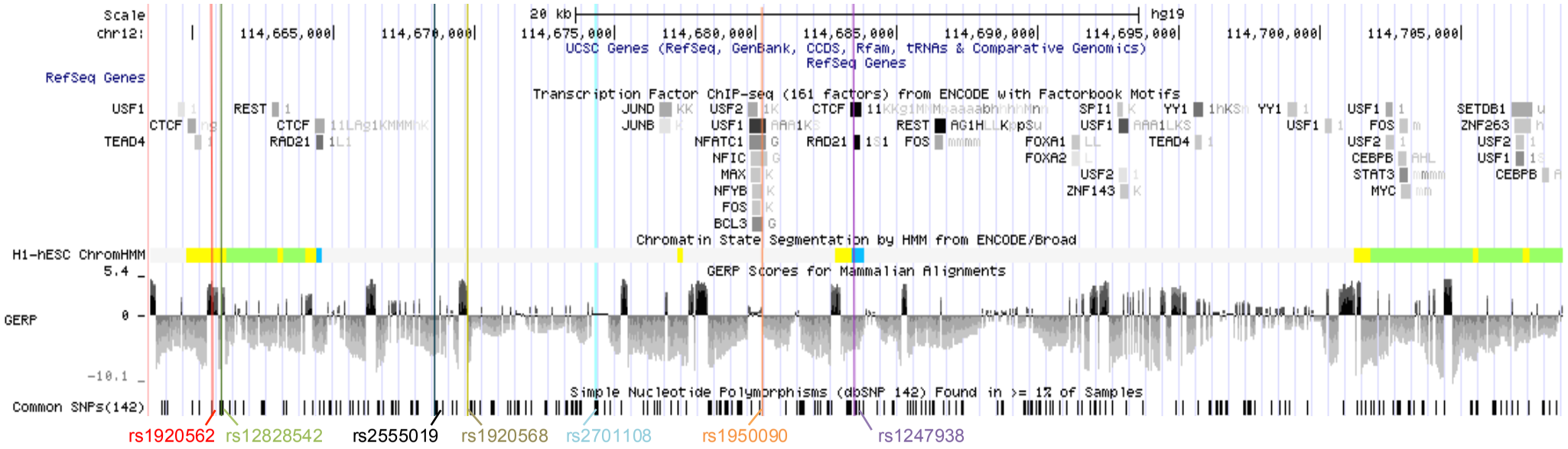


Figure 9.4: Diagrammatic representation of rs2701108 *in silico* analysis using the UCSC Genome Browser. The UCSC Genome browser (hg19 build; https://genome.ucsc.edu [5]) was interrogated for rs2701108 (indicated by a blue vertical line) and SNPs in LD (rs1920562, red vertical line; rs12828542, green vertical line; rs2555019, black vertical line; rs1920568, yellow vertical line; rs1950090, orange vertical line; rs1247938, purple vertical line). As evident from the above figure, rs2701108 does not appear to be functionally important, although HaploReg does predict binding affects on TBX5 (Table 9.5). rs1920562 and rs12828542 map to a highly conserved region, containing enhancer marks (identified by the yellow bar under H1-hESC ChromHMM). rs2555019 and rs1920568 also map to moderately conserved regions (GERP=2.13 and 2.29 respectively). rs1950090 is predicted to affect the binding of Upstream Stimulatory Factor 1 (USF1), B-Cell Lymphoma 3 (BCL3), Nuclear Transcription Factor Y, beta subunit (NFYB), Nuclear Factor 1 (NFIC) and Nuclear Factor of Activated T-cells, Cytoplasmic 1 (NFATC1). Finally, rs1247938 is predicted to alter CTCF and RAD21 binding motifs and is located within an insulator region (identified by the blue bar under H1-hESC ChromHMM).

As can be seen in Table 11.6, the highest scoring SNPs in LD with rs2701108 according to RegulomeDB were: rs1247938 (r^2=0.52), rs1920562 (r^2=0.62) and rs1950090 (r^2=0.42). CCCTC-binding factor (CTCF; a zinc finger protein involved in transcriptional repression) and the double-strand-break repair protein rad21 homolog (RAD21) binding are predicted to be affected by rs1247938 according to RegulomeDB, HaploReg and the UCSC Genome Browser (Figure 9.4 and Figure 9.5). The ability of IKAROS Family Zinc Finger 1 (IKZF1) binding is predicted to be altered by rs1920562 according to RegulomeDB, but no evidence of this was present via HaploReg or the UCSC Genome Browser. According to the UCSC Genome Browser, rs1920562 lies within a region of conservation, seen in Figure 9.4. rs1950090 is predicted to affect the binding of USF1 (Upstream stimulatory factor 1), BCL3 (B-cell lymphoma 3-encoded protein; a proto-onocogene) and NF-Y (Nuclear Factor Y; a transcription factor) according to RegulomeDB and HaploReg. The UCSC Genome Browser identified three additional binding motif changes (seen in Figure 9.4 and Figure 9.6): Nuclear Transcription Factor Y, beta subunit (NFYB), Nuclear Factor 1 (NFIC) and Nuclear Factor of Activated T-cells, Cytoplasmic 1 (NFATC1).

Sequence conservation (GERP score>2) was shown at three SNPs in LD with rs2701108: rs12828548 (r^2=0.62, GERP=2.1), rs2555019 (r^2=0.52, GERP=2.13) and rs1920568 (r^2=0.58, GERP=2.29). According to RegulomeDB, rs12828548 has “minimal binding evidence” and is predicted to alter the FOXA2 binding motif. However, according to HaploReg, the SNP is predicted to alter binding motifs of RAD21, FOXJ1 and ZFP105 (the latter two of which are transcription factors), as seen in Table 9.6. Unfortunately the UCSC Genome Browser does not support this prediction. rs12828548, according to RegulomeDB, HaploReg and the UCSC Genome Browser, lies within an enhancer region of histone modifications (Figure 9.4). rs2555019, similarly to the previous SNP, has “minimal binding evidence” according to RegulomeDB. RegulomeDB also predicts that this SNP alters a GATA binding motif, however this is not supported by HaploReg or the UCSC Genome Browser. Whilst

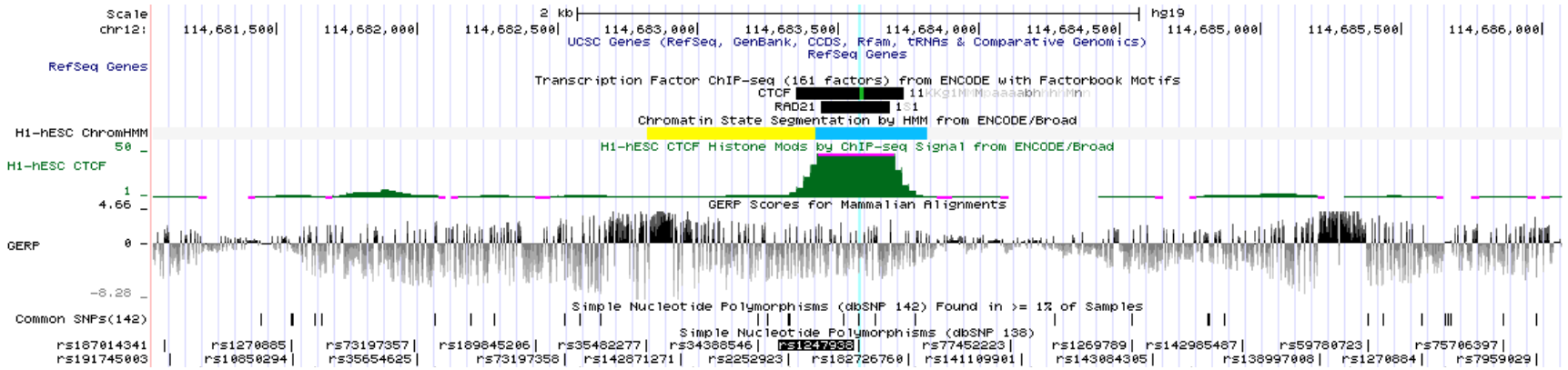


Figure 9.5: Diagrammatic representation of rs1247938 *in silico* analysis using the UCSC Genome Browser. The UCSC Genome browser (hg19 build; https://genome.ucsc.edu [5]) was interrogated for rs1247938. rs1247938 (blue vertical line) lies within an insulator region and is surrounded by histone modifications present in H1-hESC cell line. This SNP is predicted to alter CTCF and RAD21 binding motifs according to RegulomeDB, HaploReg and the UCSC Genome Browser.

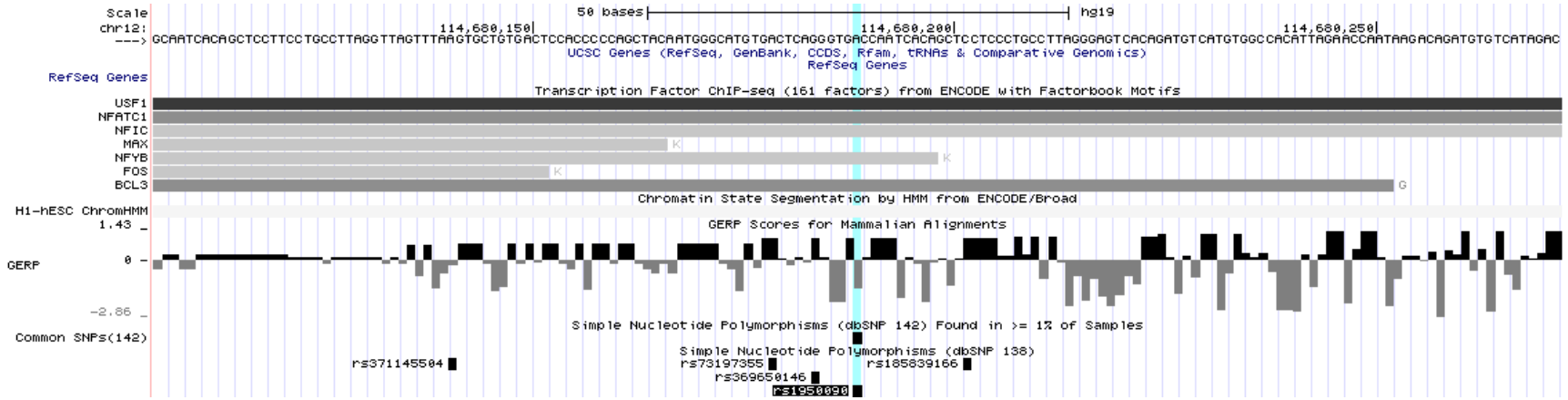


Figure 9.6: Diagrammatic representation of rs1950090 *in silico* analysis using the UCSC Genome Browser. The UCSC Genome browser (hg19 build; https://genome.ucsc.edu [5]) was interrogated for rs1950090. rs1950090 (blue vertical line) is predicted to alter USF1 (Upstream stimulatory factor 1), BCL3 (B-cell lymphoma 3), NF-Y (Nuclear Factor Y; a transcription factor), Nuclear Transcription Factor Y, beta subunit (NFYB), Nuclear Factor 1 (NFIC) and Nuclear Factor of Activated T-cells, Cytoplasmic 1 (NFATC1) binding motifs according to RegulomeDB, HaploReg and the UCSC Genome Browser.

there is no data for rs1920568 on RegulomeDB, HaploReg predicts a motif change in EWSR1/FLI1 (Ewing sarcoma breakpoint region 1/Friend leukemia integration 1 transcription factor; a chimeric fusion oncogene), ETS and STAT (Table 9.6).

An overview of this data can be found in the appendix; Table 11.6.

9.2 SNP ASSOCIATION TESTING OF THE FOUR BE VARIANTS IN EAC-ONLY CASES

BE is a known pre-cursor of EAC. To investigate whether the SNPs identified as BE risk were also linked to EAC, an independent set of 305 UK and 176 Dutch samples, reported as EAC-only were used. Four SNPs were analysed: the two newly identified SNPs (rs3072; chr2p24 and rs2701108; chr12q24) and the two previously reported in Su et al (2012) [2] (rs9257809; chr6p21 and rs9936833; chr12q24).

The first stage of the analysis compared EAC-only to Replication Phase 1 controls; no

					EAC cases *v* Controls			**BE cases *v* EAC cases**		
SNP	**CHR**	**Position**	**Alleles***	**Statistic**	**UK**	**Dutch**	**Meta**	**UK**	**Dutch**	**Meta**
rs3072	2	20741887	G/A	**OR**	1.13	1.04	1.10	1.01	1.14	1.05
				P	0.17	0.72	0.18	0.89	0.30	0.48
rs9257809	6	29464310	G/A	**OR**	1.05	0.79	0.95	0.75	0.85	0.78
				P	0.72	0.16	0.57	0.04	0.46	0.03
rs2701108	12	1.15E+08	G/A	**OR**	1.11	1.00	1.07	0.80	0.94	0.84
				P	0.23	0.98	0.34	0.01	0.63	0.02
rs9936833	16	84960619	C/T	**OR**	1.12	1.15	1.13	1.08	0.93	1.03
				P	0.19	0.25	0.08	0.38	0.58	0.66

Table 9.7: Analysis of the four Barrett's Oesophagus SNPs at genome wide significance in our studies in Oesophageal Adenocarcinoma-only cases compared with Replication Phase 1 controls and compared to Replication Phase 1 Barrett's Oesophagus cases. *Alleles are presented as minor/major. Positions are based on build 36. All results are presented in respect to the minor allele. EAC-only cases (UK N=305; Dutch N=176) had not been included in the BE case analysis. UK BE cases were from Discovery and UK Replication Phase 1 (N=2,957). Dutch BE cases were from Replication Phase 1 (N=473). Controls were from Replication Phase 1 (UK N=6,819; Dutch N=1,780). Chr, Chromosome; OR, Odds Ratio; *P*, *P* value. Note that the direction of effect for rs2701108 in the EAC-only-control analysis was the opposite of that for BE in the GWAS.

SNP was associated with EAC risk (P<0.05). The second stage of the analysis compared EAC-only cases to Replication Phase 1 BE cases; two SNPs (rs9257809 and rs2701108) were significant (P<0.05) when comparing BE to EAC cases (P_{meta}=0.03 and 0.02 respectively; Table 9.7).

9.3 REPLICATION OF BE/EAC SNPS REPORTED BY LEVINE ET AL (2013) [3]

Whilst completing this research project, a combined BE/EAC GWAS paper was published by Levine et al (2013) [3]. The paper analysed a total 922,031 SNPs in the Discovery phase and 87 of 94 SNPs with $P<1\times10^{-4}$ in the replication phase, identifying four SNPs (three loci) associated with BE/EAC risk. The four BE/EAC SNPs reported were: rs2687201, rs11789015, rs10419226, and rs10423674.

9.3.1 Replication Analysis of the Four BE/EAC SNPs

Due to an overlap in controls between the Levine study and our Discovery (the WTCCC2 controls), new UK CORGI controls [127] were employed for our Discovery set (details in Section 8.9.1).

When Levine et al (2013) [3] restricted their analysis to BE-only cases (without EAC) none of the four SNPs listed above were genome-wide significant ($P<5\times10^{-8}$), although one SNP (rs10419226 within *CRTC1*) reached $P=5.5\times10^{-8}$.

When replicating the four SNPs detailed above, one SNP (rs10423674) had already been genotyped in our Discovery set. The remaining three SNPs had to be imputed. All SNPs were analysed in the Levine Discovery, Levine Replication and our own Discovery (Number of studies = 3). rs10423674 was also already typed on the Immunochip in UKREP1 and the Dutch Replication (Number of studies = 5; Table 9.8).

Of the four BE/EAC associated SNPs, two were supported in our samples: rs2687201, located near *FOXP1*, and rs10423674, one of two SNPs located in *CRTC1* (P=0.02, OR=1.14, 95%CI 1.03-1.27 and P=0.05, OR=0.94, 95%CI 0.88-1.00 respectively). The direction of effect for rs11789015 (located in *BARX1*) was the same in both studies,

however the *P* value in our study, albeit close to the 0.05 significance criteria, was not significant; hence there appears to be limited support for this SNP (*P*=0.07, OR=0.90, 95%CI 0.81-1.01). However rs10419226, the second SNP located in *CRTC1*, was not replicated in our data (*P*=0.87, OR=1.01, 95%CI 0.91-1.11).

Meta analysis between the Levine BE/EAC data with our BE data (without EAC) showed association improvement in three of the four SNPs (rs2687201, rs11889015 and rs10423674), maintaining their genome-wide significance ($P<5\times10^{-8}$) (Table 9.8). However, association of rs10419226 worsened, although it was still genome-wide significant upon addition of our data ($P=1.17\times10^{-8}$, OR=1.14, 95%CI 1.09-1.19).

A meta-analysis between the Levine BE data (without EAC) with our BE data (without EAC) showed association improvement, with *P* values closer to the $P<5\times10^{-8}$ threshold, in three out of the four SNPs (rs2687201, rs11889015 and rs10423674). rs2687201, near *FOXP1* reached genome-wide significance ($P=4.61\times10^{-8}$, OR=1.16, 95%CI=1.10-1.23). However, the significance of rs10419226 worsened with the addition of our data ($P=2.14\times10^{-6}$, OR=1.13, 95%CI 1.08-1.20; Table 9.8).

9.3.2 Replication Analysis of the Remaining 83 BE/EAC SNPs

The remaining 83 SNPs with $P<10^{-4}$ in the BE/EAC study (Supplementary Table 3 of Levine et al (2013) [3]) were then analysed in our Discovery data to see if the inclusion of our data increased the significance of any variant. Of these, 73 were directly genotyped in our Discovery samples or were imputed with an info score of >0.95. Of the 10 remaining SNPs that could not be imputed with high quality, only one had $P<10^{-5}$ in the original Levine data; we therefore genotyped this SNP (rs11771429) using KASPar in our Discovery cases and controls. The remaining 9 SNPs were not analysed. Therefore, a total of 74 SNPs were analysed in our Discovery set.

SNP	CHR	Position	Nearby genes	Alleles[+]	Pheno-type	Levine Study		Our Study*		Meta		
						OR (95%CI)	*P*	OR (95%CI)	*P*	OR (95%CI)	*P*	N
rs2687201	3	70928930	FOXP1	T/G	BE	1.18 (1.10–1.26)	2.00×10^{-6}	1.14 (1.03-1.27)	1.18×10^{-2}	1.16 (1.10-1.23)	4.61×10^{-8}	3
					BE/EAC	1.18 (1.12–1.25)	5.47×10^{-9}			1.17 (1.11-1.23)	6.70×10^{-10}	3
rs11789015	9	96716028	BARX1	G/A	BE	0.85 (0.79–0.91)	5.08×10^{-6}	0.90 (0.81-1.01)	6.63×10^{-2}	0.86 (0.81-0.92)	1.38×10^{-6}	3
					BE/EAC	0.83 (0.79–0.88)	1.02×10^{-9}			0.85 (0.81-0.89)	1.14×10^{-10}	3
rs10419226	19	18803172	CRTC1	A/C	BE	1.19 (1.12–1.26)	5.54×10^{-8}	1.01 (0.91-1.11)	8.65×10^{-1}	1.13 (1.08-1.20)	2.14×10^{-6}	3
					BE/EAC	1.18 (1.12–1.24)	3.55×10^{-10}			1.14 (1.09-1.19)	1.17×10^{-8}	3
rs10423674	19	18817903	CRTC1	T/G	BE	0.85 (0.80–0.91)	1.92×10^{-6}	0.94 (0.88-1.00)	4.88×10^{-2}	0.89 (0.85-0.93)	2.99×10^{-7}	5
					BE/EAC	0.84 (0.80–0.89)	1.75×10^{-9}			0.88 (0.84-0.91)	4.87×10^{-11}	5

Table 9.8: Replication Analysis of four combined Barrett's Oesophagus/Oesophageal Adenocarcinoma SNPs, reported in Levine et al, in our Barrett's Oesophagus data [1]. [+]Alleles are shown as minor/major. Results shown with respect to the minor allele. *Genotypes for all SNPs were available for our discovery phase. The minimum that the meta-analysis consists of is the Levine *et al* discovery and replication phase and our discovery (N=3). rs10423674 was additionally genotyped in our UK and Dutch replication phase 1 (N=5). Note that "This study" includes the amended UK Discovery set (as described in Materials and Methods). Chr, Chromosome; BE; Barrett's Oesophagus; EAC, Oesophageal Adenocarcinoma; OR, Odds Ratio; 95%CI, 95% Confidence Intervals; *P*, *P* value; N, Number of studies.

The data was then combined by meta-analysis with the Levine BE/EAC data and our UK Discovery data. Of the 74 SNPs, six SNPs reached a *P* value $<10^{-5}$. The six SNPs were rs1497205, rs254348, rs3784262, rs6479527, rs9837992 and rs4523255. However, the SNAP database showed that rs6479527 was correlated with one of the four BE/EAC associated SNPs, rs11789015 (r^2=0.39), so is therefore unlikely to be an independent signal. Similarly, rs9837992 is in LD with rs2687201 (r^2=0.96), another of the four BE/EAC SNPs. Hence, these two SNPs were not taken forward for further replication. The remaining 4 SNPs (rs1497205, rs254348, rs3784262 and rs4523255) were genotyped in our Replication Phase samples.

All four of these SNPs were genotyped in the Levine Discovery, Levine Replication, UK Discovery, UKREP1, Dutch Replication and Extension (analysed as one sample set) and UKREP2 (number of studies=6). rs3784262 was taken forward due to the *P* value nearing the significance threshold ($P<5\times10^{-8}$), and was genotyped in the Irish Replication, UKREP3 and the Belgium Replication (number of studies=9).

Upon meta-analysis of the Levine BE/EAC data and our BE data, three out of the four SNPs (rs1497205, rs4523255 and rs3784262) showed improvement in their association. rs1497205 (located near *PARM1*) had a *P* value in the Levine data of 1.28×10^{-5} (OR=0.87, 95%CI 0.82-0.93), now with a meta *P* value of 3.68×10^{-7} (OR=0.90, 95%CI 0.86-0.94). rs4523255 (located near *MFHAS1*) had Levine *P* value of 4.15×10^{-5} (OR=1.13, 95%CI 1.07-1.20), now with a meta *P* value of 9.24×10^{-6} (OR=1.09, 95%CI 1.05-1.14). Whilst rs3784262 (located near *ALDH1A2*), with a Levine *P* value of 6.72×10^{-7} (OR=0.88, 95%CI 0.83-0.92), was now associated with BE/EAC risk with a meta *P* value of 3.72×10^{-9} (OR=0.90, 95%CI 0.87-0.93; Table 9.9).

In a BE-only meta-analysis of the two studies, two out of four SNPs (rs1497205 and rs4523255) showed association improvement, with more significant *P* values upon addition of our samples. However, neither SNP was associated with a BE-only risk ($P<5\times10^{-8}$). rs1497205 had a Levine *P* value of 2.86×10^{-5} (OR=0.86, 95%CI 0.80-0.92), and a meta *P* value of 2.57×10^{-6} (OR=0.90, 95%CI 0.86-0.94). rs4523255 had a Levine

P value of 2.46×10^{-4} (OR=1.13, 95%CI 1.06-1.21) and a meta *P* value of 2.48×10^{-5} (OR=1.09, 95%CI 1.05-1.14). The significance of the remaining two SNPs (rs254348 and rs3784262) decreased upon addition of our BE data. The Levine *P* values for rs254348 and rs3784262 were 1.15×10^{-4} (OR=0.88, 95%CI 0.83-0.94) and 3.62×10^{-7} (OR=0.85, 95%CI 0.80-0.90) respectively, with a meta *P* values of 5.49×10^{-4} (OR=0.93, 95%CI 0.89-0.97) and 1.37×10^{-6} (OR=0.91, 95%CI 0.87-0.94) respectively (Table 9.9).

SNP	CHR	Position	Nearby genes	Alleles[+]	Pheno-type	Levine Study		Our Study*		Meta		
						OR (95%CI)	*P*	OR (95%CI)	*P*	OR (95%CI)	*P*	N
rs1497205	4	76169067	PARM1, RCHY1	C/T	BE	0.86 (0.80-0.92)	2.86×10^{-5}	0.92 (0.87-0.98)	7.59×10^{-1}	0.90 (0.86-0.94)	2.57×10^{-6}	6
					BE/EAC	0.87 (0.82-0.93)	1.28×10^{-5}			0.90 (0.86-0.94)	3.68×10^{-7}	6
rs254348	16	65980789		T/C	BE	0.88 (0.83-0.94)	1.15×10^{-4}	0.95 (0.91-1.01)	8.88×10^{-2}	0.93 (0.89-0.97)	5.49×10^{-4}	6
					BE/EAC	0.89 (0.84-0.94)	1.40×10^{-5}			0.92 (0.89-0.96)	2.81×10^{-5}	6
rs3784262	15	58253106	ALDH1A2	G/A	BE	0.85 (0.80-0.90)	3.62×10^{-7}	0.93 (0.89-0.98)	5.13×10^{-3}	0.91 (0.87-0.94)	1.37×10^{-6}	9
					BE/EAC	0.88 (0.83-0.92)	6.72×10^{-7}			0.90 (0.87-0.93)	3.72×10^{-9}	9
rs4523255	8	8713038	MFHAS1	A/G	BE	1.13 (1.06-1.21)	2.46×10^{-4}	1.07 (1.01-1.12)	2.11×10^{-2}	1.09 (1.05-1.14)	2.48×10^{-5}	6
					BE/EAC	1.13 (1.07-1.20)	4.15×10^{-5}			1.09 (1.05-1.14)	9.24×10^{-6}	6

Table 9.9: Replication Analysis of four selected SNPs based on *P* value, from a total of 83 with $P<1\times10^{-4}$ from Levine *et al* in our Barrett's Oesophagus data [1]. [+]Alleles shown as minor/major. Results shown with respect to the minor allele. *Genotypes for all SNPs were available for our discovery phase. All four SNPs were genotyped in the Levine *et al* discovery and replication phase, our discovery, UK Replication 1, Dutch Replication and Extension, and UK replication phase 2 (N=6). rs3784262 was also genotyped in the Irish Replication samples, UK Replication 3 and Belgium Replication samples (N=9). Note that "This study" includes the amended UK Discovery set (as described in Methods). Chr, Chromosome; BE, Barrett's Oesophagus; EAC, Oesophageal Adenocarcinoma; OR, Odds Ratio; 95%CI, 95% Confidence Intervals; *P*, *P* value; N, Number of studies.

10. DISCUSSION

10.1 IDENTIFICATION OF NOVEL BE-ASSOCIATED SNPS

As can be seen in Table 11.1, prior to Replication Phase 3, rs3072 was already genome-wide significant, with $P_{meta}=4.0\times10^{-9}$, OR=1.15, 95%CI=1.11-1.20. However, rs2701108 was not significant, with $P_{meta}=1.4\times10^{-7}$, OR=1.12, 95%CI=1.08-1.16, prior to Replication Phase 3.

The addition of Replication Phase 3 samples (4697 cases and 5205 controls) has increased our power to detect true variations associated with BE. In our study, the inclusion of all Replication Phase 3 samples has increased the association signal of rs3072 and rs2701108. A meta-analysis of all data (with a total sample size of 10,158 BE cases and 21,062 controls) has identified two intergenic SNPs associated with BE risk ($P<5\times10^{-8}$). The first SNP, rs3072, located on chr2p24, has $P_{meta}=1.8\times10^{-11}$, OR=1.14, 95%CI=1.09-1.18. The second SNP, rs2701108, located on chr12q24, has $P_{meta}=7.5\times10^{-9}$, OR=1.11, 95%CI=1.08-1.16.

10.1.1 rs3072 Associated Genes and Theoretical Function

The flanking genes of rs3072 are Growth Differentiation Factor 7 (*GDF7*; 7.5kb downstream) and Chromosome 2 Open Reading Frame 43 (*C2ORF43*; 6.5kb downstream). *GDF7* (also known as Bone Morphogenetic Protein 12; *BMP12*) encodes the GDF7 (BMP12) protein, which is part of the Bone Morphogenetic Protein (BMP) family of Transforming Growth Factor β (TGFβ) superfamily proteins. There is evidence that this protein plays a role in the neural system and tendon/ligament development and repair, and also regulates signalling pathways that have the potential to impact on oesophageal development [157-159]. Any change in regulation or expression of this gene could, theoretically, influence the diaphragmatic tendon, which could predispose individuals to hiatal hernias. The presence of hiatal hernias increases the likelihood of reflux, a known risk factor for BE. *C2ORF43* encodes the C2ORF43 protein and has

been implicated in prostate cancer susceptibility [160, 161], although there is no direct link to BE.

Upon *in silico* functional analysis, rs3072 was found to lie within an enhancer region of histone modifications along with three other SNPs in LD (rs9306894, r^2=0.97; rs9306895, r^2=0.97; and rs7255, r^2=0.60). rs3072 was found to alter a GATA (according to HaploReg) and Regulatory Factor X3 (RFX3; according to RegulomeDB) binding motif. GATA factors are zinc finger DNA binding proteins that regulate transcription; playing a role in proliferation, differentiation and apoptosis. Altered expression of GATA4, GATA6 and Indian Hedgehog (IHH; a GATA6 target) has already been identified in BE [162]. RFX3 is a transcription factor, encoded by the *RFX3* gene, which contains a highly conserved winged helix DNA binding domain. It appears to play a role in the regulation of genes involved in ciliary assembly and motility [156] and has also been implicated in cell growth and carcinogenesis in the skin through Ras-signalling [163]. Although rs3072 does not appear to be at a conserved base or region after interrogation of the ANNOVAR database (GERP<2 and no PhastCons score), two SNPs in LD with rs3072 were (rs7255, r^2=0.60 and rs13385191, r^2=0.46). In particular, rs7255 is also predicted to be within the same enhancer region as rs3072, suggesting functional importance at this base (rs7255). eQTL studies in fibroblasts showed *C2orf43* as the most likely target of rs13385191, an intronic SNP of *C2orf43* according to RegulomeDB. C2ORF43 protein expression, determined via the Human Protein Atlas (http://www.proteinatlas.org/ENSG00000118961-C2orf43/tissue), was recorded as 'medium' and RNA expression as 14FPKM. Unfortunately there was no protein expression data for *GDF7*, although RNA expression in the oesophagus was 1FPKM (http://www.proteinatlas.org/ENSG00000143869-GDF7/tissue).

Although RegulomeDB identified *C2ORF43* as the likely target of one SNP in LD with rs3072 (rs13385191, r^2=0.46), the impact on *GDF7* cannot be discounted, due to the lack of available expression data. When reviewing previous literature, *GDF7* appears to

be the most likely target of rs3072 and/or SNPs in LD. The BMP family and TGFβ pathway have long been established as a possible mechanisms of BE, particularly in the transformation of normal oesophageal squamous cells into columnar cells [164, 165].

10.1.2 rs2701108 Associated Genes and Theoretical Function

The closest genes to the genotyped SNP, rs2701108, are T-Box 5 (*TBX5*; 117kb downstream) and RNA binding motif protein 19 (*RBM19*; 270kb upstream). Whilst the imputed, more significant SNP at this locus (rs1920562, r^2=0.62), lies 131kb downstream of *TBX5* and 256kb upstream of *RBM19*. *TBX5* encodes the TBX5 protein, a member of the conserved family of genes, which share a DNA-binding domain (T-box). *TBX5* is an important transcription factor involved in developmental regulation and its protein is involved in heart and upper limb development. Mutations within the *TBX5* gene have been associated with Holt-Oram syndrome (HOS) and Tetralogy of Fallot (TOF) syndrome, developmental disorders affecting the heart and upper limbs [166, 167]. This could in theory, be linked to BE, where abnormalities of the diaphragmatic musculature could predispose patients to hiatal hernias and acid reflux, two known BE risk factors [168, 169]. *RBM19* encodes a nucleolar protein that contains six RNA-binding motifs. The RBM19 protein is thought to be involved in ribosome biogenesis and the regulation of proliferation, differentiation and cell fate decision in the intestinal epithelium [170] and has also been implicated in a combined HOS and ulnar-mammary syndrome with *TBX5* and *TBX3* [171]. Therefore, if the expression of this gene was altered it could, theoretically, have a role in IM seen in some BE patients.

At first glance, *in silico* functional analysis of rs2701108 suggests that it is not functionally important when compared to SNPs in LD. However, its importance may lie in HaploReg's prediction of this SNP to alter a TBX5 binding motif. Whilst the imputed SNP, rs1920562, which had a stronger association upon imputation ($P_{Discovery}=1.4\times10^{-5}$, OR=1.19 compared to rs2701108: $P_{Discovery}=1.4\times10^{-3}$, OR=1.14), maps to a highly conserved base and a region containing enhancer marks. This suggests that the

imputed SNP is a more promising functional candidate, due to its conservation, instead of rs2701108.

Whilst three other SNPs were also recorded as conserved (rs12828548, r^2=0.62; rs2555019, r^2=0.52; and rs1920568, r^2=0.58), only rs12828548 lies within a region of histone modifications; suggesting functional importance for this particular SNP. However, predicted binding motif changes for rs12828548 are not consistent between RegulomeDB (FOXA2) and HaploReg (RAD21, FOXJ1 and ZFP105).

Expression analyses in RegulomeDB did not suggest the target of the chr12q24 variation, however, the Human Protein Atlas did provide protein and RNA expression data in the oesophagus. It recorded the TBX5 protein expression as 'medium' (with 30/79 tissue cell types recorded as 'medium' or 'high') and RNA expression at 3FPKM (http://www.proteinatlas.org/ENSG00000089225-TBX5/tissue). Whilst the RBM19 protein level was recorded at 'medium' (with 51/76 tissue cell types recorded as 'medium' or 'high') and RNA expression levels at 8FPKM (http://www.proteinatlas.org/ENSG00000122965-RBM19/tissue).

Whilst there is no definitive target of this variation within the selected databases, *TBX5* appears to be the most likely target, based on the current literature. In particular, diaphragmatic musculature abnormalities seen in patients with *TBX5* variations are of interest due to the possible predisposition to hiatal hernias and acid reflux.

10.1.3 Limitations

The first limitation in this section of the replication study was the conclusive definition of BE. All BE patients recruited as part of the ChOPIN clinical trial had to meet the strict Prague criteria with measurements of C1M1 or C0M2 (as detailed in section 8.1.1). It is possible that mistakes could have been made when recording these figures. In order to minimise potential errors, the CI (Professor Janusz Jankowski) checked all patient history forms, endoscopy reports and histology reports to ensure that all patients had a confirmed diagnosis of BE and the correct measurements were taken. Secondly, the

lack of conclusive expression data meant that possible gene targets could not be confidently identified.

10.2 ANALYSIS OF SNPS IN EAC CASES

Analysis of the four BE-associated SNPs rs3072 (chr2p24), rs9257809 (chr6p21) [2], rs2701108 (chr12q24) and rs9936833 (chr16q24) [2] in a small EAC sample set, showed no association (P>0.05) when comparing EAC cases to controls (P_{meta}=0.18, 0.57, 0.34 and 0.08 respectively). Two SNPs (rs9257809 and rs2701108) were significant (P<0.05) when comparing BE to EAC cases (P_{meta}=0.03 and 0.02 respectively), suggesting their involvement is associated with BE rather than EAC. However, there are limitations attached to this analysis. Firstly, the number of cases used is considerably small when analysing genome-wide significant BE SNPs in a separate disease (EAC) sample set (305 UK and 176 Dutch). Secondly, there is no guarantee that all EAC cases analysed did not also have BE, either before or as well as the presence of EAC, as BE can be asymptomatic. Therefore they may not have been true EAC-only cases. In order to improve upon this further, more EAC cases who have been under endoscopic surveillance and are known not to have had, or have, BE need to be analysed.

10.3 REPLICATION OF LEVINE ET AL (2013) [3] SNPS

10.3.1 Four BE/EAC-associated SNPs

Our replication of the four BE/EAC-associated SNPs reported in Levine et al (2013) [3] showed support for two SNPs (rs2687201 and rs10423674) and limited support for a third SNP (rs11889015). However, one of two genome-wide significant SNPs in *CRTC1* (rs10419226) was not supported. Interestingly, the SNP not supported in our data was the strongest association in the Levine et al (2013) [3] study. We also found rs2687201, near *FOXP1* to be genome-wide significance in a BE-only meta-analysis (P=4.61×10^{-8}, OR=1.16, 95%CI=1.10-1.23). *FOXP1* is a transcription factor important

in the regulation of oesophageal muscle development as described by Shu et al (2007) [172].

The limitations here were the number of cohorts each SNP was analysed in, and hence the total sample size. Three SNPs were analysed in three cohorts (Levine Discovery, Levine Replication and our Discovery), whilst one (the other SNP near *CRTC1*) was additionally genotyped in two extra cohorts (UKREP1 and Dutch). This could account for the support of one of the *CRTC1* SNPs (number of studies=5) and not the other (number of studies=3), as the likelihood of detecting the effect is reduced due to decreased statistical power. Another limitation was the different genotyping methods; imputation of three SNPs and direct genotyping of one SNP, although all three imputed SNPs had an info score >0.95.

10.3.2 Analysis of Four Selected SNPs from Levine et al (2013) [3]

All four SNPs selected were supported in our study with regards to direction of effect, however the *P* values were not significant in our study alone. In a BE-only meta-analysis of the Levine et al (2013) [3] study and this study, two out of four SNPs showed association improvement, suggesting limited association with disease; but none reached genome-wide significance. Upon meta-analysis of the Levine et al (2013) [3] BE/EAC data and our BE data (hence increasing statistical power), three of the four SNPs showed improvement in their association, suggesting limited disease association; with rs3784262 (located near *ALDH1A2*), reaching genome-wide significance; $P_{meta}=3.72\times10^{-9}$, OR=0.90, 95%CI 0.87-0.93. Aldehyde dehydrogenase 1 family, member A2 (*ALDH1A2*) encodes an enzyme that catalyzes the synthesis of retinoic acid (RA) from retinaldehyde. It is thought to be involved in many cellular processes, including alcohol metabolism [173], where a recent study has identified methylation of a CpG site near this gene to be associated with loss of control over drinking [174]. This gene has also been implicated in TOF syndrome (similar to *TBX5*) [175] and is a candidate tumour suppressor gene in prostate cancer [176].

The limitation in this section of the study was the sample size in which the SNPs were genotyped. Meta-analysis was performed on all SNPs once they were genotyped in 6 cohorts. One SNP (rs3784262) was then genotyped in a further 3 cohorts, due to the *P* value nearing the significance threshold. In order to gain a clear overview of the disease association of each SNP, all should have been genotyped in the same number of cohorts, so that the statistical power of all SNPs analysed was the same.

10.4 FUTURE RESEARCH

The following research needs to be fulfilled in order to gain a greater understanding of the disease:

1. To genotype rs1920562 (the imputed SNP at 12q24 found to be more statistically significant ($P_{Discovery}$=1.4×10^{-5}, OR=1.19) than the lead genotyped SNP rs2701108 ($P_{Discovery}$=1.4×10^{-3}, OR=1.14)) in all replication samples. To then determine, with the same statistical power, which SNP (rs2701108 or rs1920562) is highly associated with BE.
2. To obtain and analyse a greater number of pure EAC-only cases, increasing statistical power, for SNP association testing.
3. To ensure that statistical power is the same throughout, there is a need to genotype and analyse seven Levine et al (2013) [3] SNPs (the four BE/EAC associated SNPs and three of the four selected for replication) in the same number of our BE-only samples (as was done for rs3784262).

10.5 OVERVIEW

The information generated from the Levine et al (2013) [3] BE/EAC GWAS, the Su et al (2012) [2] BE GWAS and the Palles et al (2015) [1] BE Replication study allows the generation of hypotheses regarding the potential processes involved in BE. Firstly, transcription factors appear to play an important role, particularly those involved in development and structure of the thorax, diaphragm and oesophagus. The SNPs near *FOXF1* [2], *FOXP1* [3], *BARX1* [3], *GDF7* [1] and *TBX5* [1] support this hypothesis.

Secondly, the inflammatory response may play an important role, a theory supported by the SNP located within the HLA region [2]. Theoretically, the two groups of SNPs together could influence the development BE via the onset of GERD, perhaps through diaphragmatic structure (hiatal hernias), leading to an inflammatory response to the refluxed gastric acid within the oesophagus. Clinical trials provide an important source of information and sample collection for disease research. Trials investigating drug therapy also provide insight into the possible disease mechanisms; we therefore await with interest the outcome of clinical trials such as AspECT.

11. APPENDICES

Figure 11.1: Patient History Form from the ChOPIN clinical trial protocol (Version 6).

ChOPIN Investigator Reported Patient History Form

To be completed for all patients willing to participate in ChOPIN

REC REFERENCE: 06/Q1603/07

Site:	Investigator:	
Study Number: ChP	Pt Initials: First | Mid | Last	Date of birth: dd | mon | yyyy

PLEASE COMPLETE ALL SECTIONS

(To avoid duplication in entering the data for Participants in both BOSS and ChOPIN, please attach BOSS INITIAL HISTORY SHEET and complete any additional information required on this form)

Inclusion criteria – please tick box to confirm the met criteria

Aged 18 or over ☐

Patient/subject or legal representative is willing and able to give informed consent for participation in the study ☐

With any one or combination of the following (please tick all that apply):

- Endoscopically diagnosed with reflux oesophagitis, grade C or D according to Los Angeles Reflux criteria
  - endoscopy report required (and histology if available) from within 2 years of recruitment ☐
- Endoscopically and histologically diagnosed Barrett's oesophagus of C1M1 or C0M2 or greater (by Prague C&M criteria, circumferential Barrett's 1cm or greater or a 2cm or greater tongue of Barrett's)
  - endoscopy and histology reports required from within 2 years of recruitment ☐
- Histologically proven low or high grade dysplasia in Barrett's oesophagus
  - endoscopy and histology reports required from within 2 years of recruitment ☐
- Histologically proven oesophageal cancer
  - histology report confirming diagnosis required (and endoscopy report if available) ☐

SECTION 1

PATIENT DETAILS

Gender: M ☐ **F** ☐ **Weight**_____**Kg** **Height** _____**cm**

Ethnicity:

(a)	*WHITE*	*(b)*	*(BLACK or BLACK BRITISH)*	*(c)*	*ASIAN or ASIAN BRITISH*
☐	British	☐	Caribbean	☐	Indian
☐	Irish	☐	African	☐	Pakistani
				☐	Bangladeshi
☐	Any other White background	☐	Any other Black background	☐	Any other Asian background
	please write in below		*please write in below*		*please write in below*
	……		……..		……...
(d)	*MIXED*	*(e)*	*CHINESE or OTHER ETHNIC GROUP*		
☐	White and Black Caribbean	☐	Chinese	☐	Patient refusal
☐	White and Black African	☐	Any other Mixed background		
☐	White and Asian		*please write in opposite*		
☐	Any other Mixed background				
	please write in below				
	……				

SECTION 2

ALCOHOL INTAKE

Yes ☐ No ☐

If yes, please specify amount/week: __ __ units

Please calculate amount of pure alcohol consumed per week, using the conversion table below

1 pt of beer	2 units
Spirit (25 ml)	1 unit
Spirit (35 ml)	1.5 units
Glass of wine (125 ml)	1.5 units

SMOKING HISTORY

current smoker ☐ previous smoker ☐ never smoked ☐

For current and previous smokers
Years of smoking: …… Number per day: ……

BARRETT'S HISTORY

Duration of reflux symptoms: _____Year(s)

Date of endoscopy confirming Barrett's oesophagus Date : dd | Mon | yyyy

Date of most recent endoscopy if different from above Date : dd | Mon | yyyy

Regular surveillance for Barretts oesophagus prior to trial entry No ☐ Yes ☐

ENDOSCOPY FINDINGS AT STUDY ENTRY (OR WITHIN 2 YEARS)

Length of circumferential Columnar Lining only) **C** |___|___|. |___|cm (Barrett;'s

Length of Tongues of Columnar Lining (maximal extent) only) **M** |___|___|. |___|cm (Barrett's

Was a hiatal hernia present? No☐ Yes☐ Size:.....

Was intestinal metaplasia present? No☐ Yes☐

Was indefinite dysplasia present? No☐ Yes☐

Was low grade dysplasia present? No☐ Yes☐

Was high grade dysplasia present? No☐ Yes☐

Mucosal break(s) / Oesophagitis No☐ Yes☐

(if yes - Los Angeles Classification grade A☐ B☐ C☐ D☐)

Was oesophageal adenocarcinoma present? No ☐ Yes ☐

Was oesophageal squamous cell cancer present? No ☐ Yes ☐

Was oesophageal poorly differentiated epithelial malignancy present? No ☐ Yes ☐

Was a junctional tumour present? No ☐ Yes-Siewert type I ☐ type II ☐

HELICOBACTER TEST

Not taken ☐ If taken, was the result:

Positive ☐
Positive and Eradicated ☐
Negative ☐
Not known ☐

DOCUMENTED HISTORY OF VASCULAR DISEASE

Does the patient have a history of any of the following (please tick)**:**	**No**	**Yes**
Myocardial infarction:	☐	☐
Angina (physician diagnosed):	☐	☐
Coronary intervention (e.g. prior bypass surgery, coronary angioplasty or stent):	☐	☐
Carotid artery stenosis on ultrasound or angiography:	☐	☐
Cerebrovascular accident:	☐	☐
Transient ischaemic attack:	☐	☐
Peripheral vascular disease (e.g. history of claudication, prior peripheral vascular disease):	☐	☐
Diabetes mellitus:	☐	☐
Hypertension:	☐	☐
Hyperlipidaemia and high cholesterol	☐	☐
(LDL >130 mg/dl (or >3.4 mmol/L) and/or HDL <40 mg/dl (or <1.0mmol/L)):	☐	☐

SECTION 3

FAMILY HISTORY-digestive tract related conditions only **No** ☐ **Yes** ☐
(please give details, below)

Relation to patient (brother, sister, parent, or child)	Heartburn	Barrett's	Age at diagnosis	Oesophageal Cancer (type)	Age at diagnosis

UPPER GI SURGERY

Previous upper gastrointestinal surgery: **No** ☐ **Yes** ☐ (please give details, below)

Type of Surgery	Date of Surgery	Outcome
	Mon | yyyy	
	Mon | yyyy	
	Mon | yyyy	

CURRENT MEDICATION- digestive tract related medication and long term (≥ 3 months) NSAID/ aspirin use only

Drug (brand name)	Dosage Dose (Including units)	Frequency	Route e.g. IV	**Form** e.g. tablet	Indication	Start Date
						Mon | yyyy
						Mon | yyyy
						Mon | yyyy

						Mon | yyyy
						Mon | yyyy

FORM COMPLETED BY: _____ (print name)

SIGNATURE _____ Date: DD | MON | YYYY

Site: Hospital name
Investigator: The consultant responsible for the care of the patient.
Trial number: This is the unique number that identifies this patient in the ChOPIN study pre-determined at the ChOPIN study office and will be supplied to the trial site
Date of Birth: The patient's date of birth must be written in the following format dd/mon/yyyy, e.g. 01/FEB/1977.
Pt Initials: Record the patient's first, middle and last initial in the spaces provided. If the patient has no middle initial, please record a dash
Completed by Print name clearly, sign and provide the date when the form is completed in the correct format (see above).

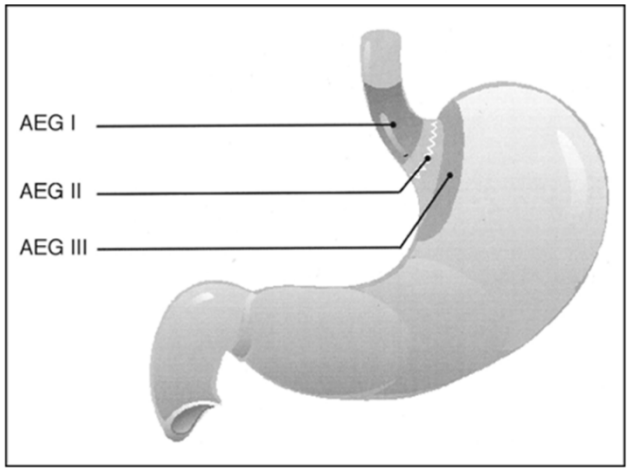
Please return the completed form together with endoscopy and pathology report forms (participant identifiable information deleted) to:

ChOPIN study Office
The John Bull Building
Plymouth University Peninsula School of Medicine and Dentistry
Tamar Science Park
Plymouth
Devon, PL6 8BU

Tel Number: 01752 437402
Fax Number: 01752 517842

Forms must be completed in black ball-point pen
Cross out errors with a single stroke, insert the correction and initial & date the change.
Correction fluid and /or sticky labels must not be use

Figure 1: Siewert Classification (for adenocarcinoma of the oesophagogastric junction (AEG)



von Rahden B H et al. JCO 2005;23:874-879

			Sequenom iPLEX			Phase 2 Meta		Discovery + Phases 1 & 2 Meta		
SNP	**Chr**	**Position**	**Design Success**	**Call Rate**	**Reason for Inclusion**	**OR (95%CI)**	***P***	**OR (95% CI)**	***P***	**N**
rs41341748	8	16056965	YES	1	*MSR1* candidate variant	1.09 (0.64-1.54)	0.701	1.09 (0.64-1.54)	0.701	1
rs3072	2	20741887	YES	1	meta $P<1\times10^{-4}$	1.13 (1.04-1.23)	0.007	1.15 (1.11-1.2)	4.04×10^{-9}	5
rs6751791	2	35435501	YES	0.998	meta $P<1\times10^{-4}$	0.94 (0.85-1.03)	0.164	0.88 (0.84-0.93)	1.82×10^{-7}	5
rs10083033	12	1.04E+08	YES	1	meta $P<1\times10^{-4}$	1.00 (0.91-1.09)	0.943	0.90 (0.85-0.94)	5.82×10^{-6}	5
rs189247	15	95387634	YES	0.999	meta $P<1\times10^{-4}$	0.91 (0.82-1.00)	0.049	0.88 (0.83-0.92)	5.00×10^{-8}	5
rs2043633	16	5759275	YES	0.998	meta $P<1\times10^{-4}$	0.88 (0.79-0.97)	0.005	0.87 (0.82-0.92)	6.02×10^{-9}	5
rs3923500	13	85657509	YES	1	meta $P<1\times10^{-4}$	0.94 (0.80-1.07)	0.336	1.13 (1.07-1.19)	5.96×10^{-5}	5
rs2731672	5	1.77E+08	YES	0.999	meta $P<1\times10^{-4}$	0.91 (0.81-1.02)	0.082	0.88 (0.82-0.93)	1.07×10^{-6}	5
rs2701108	12	1.13E+08	YES	0.993	meta $P<1\times10^{-4}$	0.89 (0.80-0.98)	0.011	0.88 (0.83-0.93)	1.39×10^{-7}	5
rs3734960	7	1.54E+08	YES	0.999	meta $P<1\times10^{-4}$	0.99(0.89-1.09)	0.814	1.11 (1.05-1.16)	1.35×10^{-4}	5
rs6921589	6	25530348	YES	1	meta $P<1\times10^{-4}$	1.01 (0.88-1.14)	0.88	1.15 (1.08-1.21)	6.48×10^{-5}	5
rs2218260	15	56001502	YES	0.935	meta $P<1\times10^{-4}$	1.05 (0.96-1.14)	0.313	1.11 (1.06-1.16)	1.11×10^{-5}	5
rs4792891	17	41329294	YES	0.995	meta $P<1\times10^{-4}$	1.00 (0.91-1.09)	0.971	0.91 (0.86-0.96)	1.44×10^{-4}	5

rs6480314	10	69637453	YES	0.989	discovery $P<1\times10^{-4}$, not on immunochip	1.03 (0.90-1.16)	0.659	1.16 (1.08-1.24)	2.02×10^{-4}	2
rs9404438	6	1.04E+08	YES	0.989	discovery $P<1\times10^{-4}$, not on immunochip	0.99 (0.88-1.09)	0.832	0.89 (0.82-0.96)	5.08×10^{-4}	2
rs7825744	8	72180729	YES	1	meta $P<1\times10^{-4}$	1.08 (0.93-1.23)	0.316	1.18 (1.1-1.26)	2.43×10^{-5}	5
rs1473857	3	1.03E+08	YES	1	meta $P<1\times10^{-4}$	0.98 (0.87-1.10)	0.788	1.11 (1.05-1.17)	4.43×10^{-4}	5
rs4674923	2	2.25E+08	YES	0.999	discovery $P<1\times10^{-4}$, not on immunochip	1.08 (0.98-1.17)	0.141	1.08 (1.02-1.14)	1.24×10^{-2}	2
rs2319335	3	1.49E+08	YES	0.998	discovery $P<1\times10^{-4}$, not on immunochip	0.89 (0.78-1.00)	0.036	1.07 (1.01-1.14)	4.02×10^{-2}	2
rs9074	20	44122072	YES	1	meta $P<1\times10^{-4}$	1.02 (0.92-1.12)	0.676	0.91 (0.86-0.97)	6.90×10^{-4}	5
rs6032951	20	10717074	YES	<90%, *	meta $P<1\times10^{-4}$	1.06 (0.95-1.18)	0.295	1.13 (1.07-1.19)	3.16×10^{-5}	5
rs4742902	9	1.06E+08	YES	1	meta $P<1\times10^{-4}$	1.01 (0.91-1.11)	0.815	0.91 (0.86-0.96)	5.12×10^{-4}	5
rs2425752	20	44135527	YES	0.936	meta $P<1\times10^{-4}$	0.97 (0.86-1.07)	0.528	0.90 (0.85-0.95)	7.94×10^{-5}	5
rs1158793	4	96223364	YES	0.962	discovery $P<1\times10^{-4}$, not on immunochip	1.08 (0.97-1.19)	0.171	0.87 (0.8-0.94)	3.43×10^{-5}	2
rs1357111	7	45651069	YES	1	discovery $P<1\times10^{-4}$, not on immunochip	0.98 (0.82-1.14)	0.796	0.86 (0.76-0.95)	1.65×10^{-3}	2
rs1333916	9	1.22E+08	YES	1	meta $P<1\times10^{-4}$	1.03 (0.93-1.13)	0.582	0.92 (0.87-0.97)	1.03×10^{-3}	5
rs13087427	3	1.18E+08	YES	0.995	meta $P<1\times10^{-4}$	1.02 (0.88-1.16)	0.776	0.88 (0.81-0.96)	7.29×10^{-4}	5

rs2854925	9	1.34E+08	YES	0.988	discovery $P<1\times10^{-4}$, not on immunochip	1.02 (0.92-1.11)	0.759	1.10 (1.04-1.17)	1.65×10^{-3}	2
rs958889	11	1.28E+08	YES	1	meta $P<1\times10^{-4}$	0.99 (0.88-1.09)	0.778	0.91 (0.85-0.96)	2.84×10^{-4}	5
rs12985909	19	18300383	YES	0.999	meta $P<1\times10^{-4}$	1.11 (1.02-1.20)	0.027	1.11 (1.07-1.16)	4.20×10^{-6}	5
rs12156009	8	11322629	YES	0.996	discovery $P<1\times10^{-4}$, not on immunochip	0.97 (0.88-1.06)	0.444	0.91 (0.87-0.96)	1.18×10^{-4}	5
rs4775330	15	59022174	YES	0.998	meta $P<1\times10^{-4}$	0.98 (0.88-1.07)	0.641	1.08 (1.03-1.13)	1.24×10^{-3}	5
rs563198	9	77521795	YES	0.999	meta $P<1\times10^{-4}$	0.94 (0.85-1.04)	0.245	0.90 (0.85-0.95)	5.38×10^{-5}	5
rs6926118	6	3320615	NO	*	discovery $P<1\times10^{-4}$, not on immunochip	0.98 (0.82-1.14)	0.816	0.86 (0.76-0.95)	1.30×10^{-3}	3
rs7042370	9	12775073	YES	0.999	discovery $P<1\times10^{-4}$, not on immunochip	1.03 (0.93-1.12)	0.619	0.90 (0.84-0.96)	4.84×10^{-4}	2
rs4858738	3	19826462	YES	0.998	discovery $P<1\times10^{-4}$, not on immunochip	1.04 (0.93-1.15)	0.501	0.88 (0.81-0.95)	2.55×10^{-4}	2
rs1935671	20	8331096	YES	0.863	meta $P<1\times10^{-4}$	1.00 (0.89-1.10)	0.935	1.10 (1.04-1.16)	8.66×10^{-4}	4
rs2902637	10	1.06E+08	NO	*	meta $P<1\times10^{-4}$	1.01 (0.90-1.11)	0.905	1.10 (1.05-1.16)	5.71×10^{-4}	5
rs10903038	1	24363645	YES	0.999	discovery $P<1\times10^{-4}$, not on immunochip	0.98 (0.87-1.08)	0.656	1.12 (1.05-1.18)	7.36×10^{-4}	2
rs851727	7	1.47E+08	YES	1	meta $P<1\times10^{-4}$	1.09 (0.87-1.30)	0.444	0.85 (0.74-0.96)	2.76×10^{-3}	5
rs9379897	6	26709505	YES	1	meta $P<1\times10^{-4}$	0.86 (0.73-0.99)	0.022	0.86 (0.79-0.92)	7.69×10^{-6}	5

rs9524596	13	94166840	YES	0.999	meta $P<1\times10^{-4}$	0.94 (0.85-1.03)	0.156	0.91 (0.86-0.96)	5.69×10^{-5}	5
rs6903535	6	28525201	YES	0.999	meta $P<1\times10^{-4}$	0.98 (0.89-1.08)	0.731	0.92 (0.87-0.97)	5.42×10^{-4}	5
rs6903130	6	32840188	NO	NA	meta $P<1\times10^{-4}$					
rs292808	1	33819233	NO	NA	$P<1\times10^{-4}$ in sex-differentiated analysis of discovery					
rs10210285	2	1.43E+08	YES	0.998	$P<1\times10^{-4}$ in sex-differentiated analysis of discovery	1.06 (0.87-1.25)	0.538	1.14 (1.02-1.26)	3.08×10^{-2}	2
rs7609738	3	1.83E+08	YES	0.999	$P<1\times10^{-4}$ in sex-differentiated analysis of discovery	1.06 (0.96-1.16)	0.251	1.05 (0.99-1.11)	9.39×10^{-2}	2
rs7816766	8	4523552	YES	0.991	$P<1\times10^{-4}$ in sex-differentiated analysis of discovery	0.88 (0.78-0.97)	0.008	1.12 (1.06-1.18)	2.48×10^{-4}	2
rs719527	11	1.03E+08	YES	0.998	$P<1\times10^{-4}$ in sex-differentiated analysis of discovery	0.99 (0.88-1.10)	0.872	1.08 (1.01-1.15)	2.33×10^{-2}	2
rs2715425	15	97278619	YES	0.998	*IGF1R* candidate variant	0.97 (0.86-1.09)	0.647	0.97 (0.86-1.09)	0.647	1
rs6898743	5	42638249	YES	0.999	*GHR* candidate variant	1.03 (0.91-1.15)	0.611	1.03 (0.91-1.15)	0.611	1
rs1325190	1	1.98E+08	YES	1	meta $P<1\times10^{-4}$	1.05 (0.95-1.15)	0.369	0.92 (0.87-0.97)	1.82×10^{-3}	5

rs7255	2	20742301	YES	0.901	meta P<1×10^{-4}	0.87 (0.78-0.96)	0.003	0.87 (0.83-0.92)	1.09×10^{-8}	5
rs13385191	2	20751746	YES	0.999	meta P<1×10^{-4}	1.05 (0.94-1.15)	0.375	1.11 (1.06-1.17)	1.06×10^{-4}	5
rs340620	2	20787591	YES	0.982	meta P<1×10^{-4}	0.94 (0.84-1.03)	0.175	0.90 (0.85-0.95)	1.52×10^{-5}	5
rs6727683	2	35429477	NO	NA	meta P<1×10^{-4}					
rs12993283	2	35445909	YES	0.999	meta P<1×10^{-4}	0.93 (0.84-1.02)	0.121	0.89 (0.84-0.93)	2.15×10^{-7}	5
rs819848	3	1.57E+08	YES	0.994	meta P<1×10^{-4}	1.02 (0.92-1.11)	0.755	1.10 (1.05-1.15)	3.75×10^{-4}	5
rs9824398	3	1.88E+08	YES	1	meta P<1×10^{-4}	1.06 (0.97-1.15)	0.241	1.10 (1.05-1.15)	5.65×10^{-5}	5
rs9879899	3	1.88E+08	YES	0.993	meta P<1×10^{-4}	0.95 (0.86-1.05)	0.321	0.91 (0.86-0.96)	6.59×10^{-5}	5
rs13157599	5	1.23E+08	YES	1	meta P<1×10^{-4}	0.92 (0.72-1.13)	0.461	1.19 (1.08-1.29)	1.43×10^{-3}	5
rs35936561	7	1.56E+08	YES	0.998	meta P<1×10^{-4}	0.98 (0.86-1.09)	0.682	0.91 (0.86-0.96)	1.01×10^{-4}	5
rs7836059	8	11309574	NO	NA	meta P<1×10^{-4}					
rs2898290	8	11471318	YES	0.989	meta P<1×10^{-4}	0.92 (0.83-1.01)	0.083	0.90 (0.86-0.95)	1.42×10^{-5}	5
rs12677326	8	11476634	NO	NA	meta P<1×10^{-4}					
rs13267835	8	11542328	YES	0.998	meta P<1×10^{-4}	0.97 (0.83-1.10)	0.615	0.87 (0.80-0.94)	9.11×10^{-5}	5
rs13273672	8	11649790	YES	0.999	meta P<1×10^{-4}	0.98 (0.89-1.08)	0.739	1.09 (1.04-1.14)	6.90×10^{-4}	5

rs8180912	8	11677400	YES	0.999	meta P<1×10^{-4}	1.00 (0.89-1.12)	0.988	0.90 (0.84-0.96)	3.73×10^{-4}	5
rs7895043	10	71016742	YES	0.999	meta P<1×10^{-4}	1.04 (0.95-1.13)	0.428	1.09 (1.05-1.14)	1.52×10^{-4}	5
rs1265496	12	1.13E+08	NO	NA	meta P<1×10^{-4}					
rs12903220	15	59040471	NO	NA	meta P<1×10^{-4}					
rs7168393	15	95353461	YES	0.999	meta P<1×10^{-4}	1.01 (0.85-1.17)	0.866	1.14 (1.07-1.21)	1.96×10^{-4}	5
rs7173314	15	95353711	NO	NA	meta P<1×10^{-4}					
rs2535483	15	95357916	NO	NA	meta P<1×10^{-4}					
rs9941024	16	5734313	YES	1	meta P<1×10^{-4}	0.90 (0.81-0.99)	0.03	0.89 (0.84-0.94)	6.72×10^{-7}	5
rs7200175	16	5742063	NO	NA	meta P<1×10^{-4}					
rs11866983	16	5743925	YES	0.997	meta P<1×10^{-4}	1.09 (0.99-1.19)	0.088	1.12 (1.07-1.17)	4.93×10^{-6}	5
rs9936833	16	84960619	YES	0.997	meta P<1×10^{-4}	1.02 (0.91-1.14)	0.669	1.15 (1.1-1.20)	4.68×10^{-8}	5
rs1532167	16	84961705	NO	NA	meta P<1×10^{-4}					
rs7187365	16	85069416	NO	NA	meta P<1×10^{-4}					
rs889592	16	85076757	YES	1	meta P<1×10^{-4}	0.99 (0.87-1.11)	0.921	0.90 (0.83-0.96)	3.00×10^{-4}	5
rs4792827	17	41487141	NO	NA	meta P<1×10^{-4}					

rs6040146	20	10701237	NO	NA	meta P<1×10^{-4}

Table 11.1: SNPs prioritised for further genotyping in Replication Phase 2 samples (Irish cohort of 245 cases and 473 controls and a UK cohort of 1,765 cases and 1,586 controls) listed in order of priority, based on Discovery and Replication Phase 1 meta *P* value. *SNPs in top 40 that couldn't be genotyped by sequenom were genotyped by KASPar. KASPar call rates were all > 95%. SNPs highlighted in grey were excluded because of call rates<95%, SNPs highlighted in blue failed at the design stage of the iPLEX. SNPs were selected based on the following criteria: (i) meta $P<10^{-4}$ = $P_{association}<10^{-4}$ in combined Discovery and Replication Phase 1 analysis, as described in Su et al (2012) [2] (N=63); (ii) discovery $P<10^{-4,}$ not on immunochip = $P_{association}<10^{-4}$ in Discovery Phase, but not included in the Immunochip content (N=12); (iii) $P<10^{-4}$ in sex-differentiated analysis of discovery =$P_{association}<1 \times 10^{-4}$ in a sex-stratified analysis of the Discovery phase (N=5); (iv) candidate variant = candidate polymorphisms that had previously been reported to be associated with BE and were not well tagged by the Discovery Phase or Immunochip arrays. Note that rs41341748 was also typed in UK Replication Phase 3 (OR=1.07, 95%CI 0.70-1.43, P_{meta}=0.79). Phase 2 Meta consisted of Irish and UKREP2 samples. Discovery + Phases 1 & 2 Meta consisted of Discovery, UKREP1, Dutch Replication, UKREP2 and Irish samples. Note that the Dutch Replication did not include the Dutch Extension at this stage (the Dutch Extension was only genotyped for the 7 SNPs taken through to Replication Phase 3). Chr, Chromosome; OR, Odds Ratio; *P*, P-value.

SNP	Primer	Sequence 5'-3'	N PCR Cycles
rs3072	Allele A	GAAGGTGACCAAGTTCATGCTCGTTTCCTTAAAATCACAGTACTGGA	36
	Allele G	GAAGGTGACCAAGTTCATGCTCGTTTCCTTAAAATCACAGTACTGGG	
	Common 1	ATGACGAGTGGCTGCAGCTTAGAAA	
	Common 2	TGGCTGCAGCTTAGAAAGCAAATTTCAT	
rs6751791	Allele A	GAAGGTGACCAAGTTCATGCTCAATGGAGTAAATTGCTAGCAAACTCAA	26
	Allele G	GAAGGTCGGAGTCAACGGATTAATGGAGTAAATTGCTAGCAAACTCAG	
	Common 1	TTGGAGCTCTCAGATTTTAATCCTGCT	
	Common 2	GAACTGCGTTGGAGCTCTCAGATTT	
rs2731672	Allele A	GAAGGTGACCAAGTTCATGCTCCAGGCTCATTTGTTAGGAATGTGA	26
	Allele G	GAAGGTCGGAGTCAACGGATTCAGGCTCATTTGTTAGGAATGTGG	
	Common 1	GGAATTATAAAGCTAGAGGCCTTCTCTTT	
	Common 2	AGGCCTTCTCTTTCCATGGAGGTT	
rs2701108	Allele A	GAAGGTGACCAAGTTCATGCTGGCAGGAGAAAATGTGTACTCTCAT	26
	Allele G	GAAGGTCGGAGTCAACGGATTGCAGGAGAAAATGTGTACTCTCAC	
	Common 1	CCTCCCTGCCCTGCCTCCTA	
	Common 2	CTGCCTCCTAGGTGACTCTGGAA	
rs189247	Allele A	GAAGGTGACCAAGTTCATGCTCAGATGCCCATCAGAAAACCCAA	26
	Allele G	GAAGGTGACCAAGTTCATGCTCAGATGCCCATCAGAAAACCCAG	
	Common 1	CCTGACACTCCACAGATGGAAACAA	
	Common 2	CAGATGGAAACAAACCTGGGAACCTT	
rs2043633	Allele A	GAAGGTCGGAGTCAACGGATTGACCTTATTTGGAGACAGGGTGTT	26
	Allele C	GAAGGTCGGAGTCAACGGATTGACCTTATTTGGAGACAGGGTGTG	
	Common 1	CCACCCTAATGACCTGATTTTTACTTGAT	
	Common 2	TGACCTGATTTTTACTTGATTAGCTCCATA	
rs12985909	Allele A	GAAGGTCGGAGTCAACGGATTGGTGGCCAATCAGCGTTCCT	26
	Allele G	GAAGGTCGGAGTCAACGGATTGGTGGCCAATCAGCGTTCCC	
	Common 1	GTGCTCTTTCTTTAACCAATCCCTGTA	
	Common 2	CAATCCCTGTAGCCAGGGGGAT	

Table 11.2: Primers used for the seven SNPs taken through to UKREP3. Green indicates primer used after optimization

SNP	Primer	Sequence 5'-3'	N PCR Cycles
rs1497205	Allele A	GAAGGTGACCAAGTTCATGCTAGCCTATCGTCCCTCTGGTGT	26
	Allele C	GAAGGTCGGAGTCAACGGATTGCCTATCGTCCCTCTGGTGC	
	Common 1	CCATTATGAAGCACAGAGCACCTGA	
	Common 2	CACAGAGCACCTGAGCCTATAAACAA	
rs254348	Allele A	GAAGGTGACCAAGTTCATGCTACAAGAAGGAGAGGATAAGGAGACA	36
	Allele C	GAAGGTCGGAGTCAACGGATTAAGAAGGAGAGGATAAGGAGACG	
	Common 1	CATGAGGGTCAGGCTGGGGAAT	
	Common 2	CAGGCTGGGGAATGACAGAGGAT	
rs3784262	Allele A	GAAGGTGACCAAGTTCATGCTCTGAGCACATAATCTGACTGGCAT	26
	Allele G	GAAGGTCGGAGTCAACGGATTGAGCACATAATCTGACTGGCAC	
	Common 1	CCTCTGGAGGAAAGGAATTTAAAAATAAAA	
	Common 2	CAATTTTTTCCTCTGGAGGAAAGGAATTTA	
rs4523255	Allele C	GAAGGTGACCAAGTTCATGCTCATGAATCAGGATTAAAGAATCTGTGG	36
	Allele T	GAAGGTCGGAGTCAACGGATTCATGAATCAGGATTAAAGAATCTGTGA	
	Common 1	ACTAATGTCTCCATCCTGCACAATGAAA	
	Common 2	CCATCCTGCACAATGAAAATGAACGTTAT	
rs11771429	Allele C	GAAGGTGACCAAGTTCATGCTCAAAGTTTATCGTAAAACTACAGGAGG	36
	Allele T	GAAGGTCGGAGTCAACGGATTACAAAGTTTATCGTAAAACTACAGGAGA	
	Common 1	AAAGCCAAAGGGCCTTGATAACATTAGTT	
	Common 2	GGGCCTTGATAACATTAGTTTGGAAGATT	

Table 11.3: Primers used for KASP genotyping of the Levine SNPs. Four selected Levine et al (2013) [3] SNPs taken through to replication phases and the one Levine et al (2013) [3] SNP (rs11771429) genotyped in our Discovery Phase using KASPar. Green indicates primer used after optimization.

SNP	Chr	Position	Alleles *	Statistics	UK Discovery	Dutch	UKREP1	Irish	UKREP2	UKREP3	Belgian	BEACON	Meta-analysis
				OR	**1.23**	**1.20**	**1.06**	**1.02**	**1.15**	**1.18**	**0.95**	**1.11**	**1.14**
rs3072	**2**	**20878406**	**G/A**	**95%CI**	**1.14-1.33**	**1.05-1.38**	**0.97-1.16**	**0.81-1.30**	**1.04-1.28**	**1.04-1.34**	**0.79-1.14**	**1.03-1.19**	**1.09-1.18**
				P	**2.64×10^{-7}**	**8.77×10^{-3}**	**2.22×10^{-1}**	**8.47×10^{-1}**	**4.79×10^{-3}**	**1.18×10^{-2}**	**5.58×10^{-1}**	**6.64×10^{-3}**	**1.75×10^{-11}**
				OR	1.15	1.10	1.18	1.30	1.03	0.98	0.96	1.00	1.08
rs6751791	2	35581997	A/G	**95%CI**	1.06-1.23	0.96-1.27	1.08-1.29	1.04-1.63	0.93-1.13	0.83-1.16	0.80-1.14	0.93-1.07	1.04-1.12
				P	5.03×10^{-4}	1.51×10^{-1}	3.05×10^{-4}	2.22×10^{-2}	5.97×10^{-1}	8.04×10^{-1}	6.29×10^{-1}	8.92×10^{-1}	7.65×10^{-5}
				OR	1.18	1.10	1.15	1.05	1.10	1.02	0.90	0.95	1.07
rs2731672	5	176842474	A/G	**95%CI**	1.09-1.28	0.94-1.29	1.04-1.28	0.82-1.35	0.99-1.23	0.85-1.24	0.73-1.10	0.88-1.03	1.03-1.12
				P	1.64×10^{-4}	2.18×10^{-1}	8.03×10^{-3}	6.82×10^{-1}	8.51×10^{-2}	8.13×10^{-1}	2.98×10^{-1}	2.09×10^{-1}	1.66×10^{-3}
				OR	**0.88**	**0.91**	**0.86**	**0.71**	**0.93**	**0.97**	**0.95**	**0.91**	**0.90**
rs2701108	**12**	**114674261**	**G/A**	**95%CI**	**0.81-0.95**	**0.79-1.04**	**0.78-0.94**	**0.57-0.89**	**0.84-1.03**	**0.86-1.10**	**0.80-1.14**	**0.85-0.98**	**0.86-0.93**
				P	**1.00×10^{-3}**	**1.62×10^{-1}**	**9.76×10^{-4}**	**2.86×10^{-2}**	**1.46×10^{-1}**	**6.69×10^{-1}**	**6.04×10^{-1}**	**1.43×10^{-2}**	**7.48×10^{-9}**
				OR	1.18	1.23	1.10	1.03	1.11	1.07	1.00	1.04	1.10
rs189247	15	97586630	A/G	**95%CI**	1.09-1.27	1.06-1.41	1.00-1.21	0.81-1.30	1.00-1.23	0.94-1.22	0.84-1.19	0.97-1.12	1.06-1.14
				P	5.67×10^{-5}	5.00×10^{-3}	4.36×10^{-2}	8.21×10^{-1}	4.02×10^{-2}	2.80×10^{-1}	9.89×10^{-1}	3.10×10^{-1}	3.55×10^{-7}
				OR	0.85	0.84	0.90	0.87	0.88	1.03	0.84	1.01	0.92
rs2043633	16	5819274	C/A	**95%CI**	0.79-0.92	0.74-0.97	0.82-0.98	0.70-1.09	0.80-0.97	0.90-1.17	0.70-1.00	0.94-1.09	0.88-0.95
				P	6.04×10^{-5}	1.36×10^{-2}	2.05×10^{-2}	2.28×10^{-1}	1.21×10^{-2}	6.83×10^{-1}	5.59×10^{-2}	7.87×10^{-1}	2.25×10^{-6}
				OR	1.12	1.14	1.11	1.09	1.11	1.03	1.14	1.07	1.10
rs12985909	19	18439383	G/A	**95%CI**	1.04-1.21	0.99-1.30	1.02-1.22	0.87-1.38	1.01-1.22	0.91-1.16	0.95-1.37	1.00-1.15	1.06-1.14
				P	2.94×10^{-3}	6.06×10^{-2}	1.87×10^{-2}	4.44×10^{-1}	3.80×10^{-2}	6.47×10^{-1}	1.51×10^{-1}	5.20×10^{-2}	3.28×10^{-7}

Table 11.4: Cohort breakdown for the seven selected SNPs taken into Replication Phase 3. *Alleles shown as minor/major. All results are presented with respect to the minor allele. rs6751791 was not genotyped in BEACON. Data presented are for a proxy SNP: rs7598399; r^2= 1. Also, rs189247 was not genotyped in BEACON, but was imputed from 4 genotyped SNPs (rs991757, rs2670927, rs2670930 and rs234540). Imputation accuracy using this strategy was 98.2%, confirmed by imputing samples for which genotypes from sequence data were available for all 5 SNPs and checking concordance of imputed and sequenced genotypes. Dutch cohort consists of Dutch Replication (Phase 1 replication) and Dutch Extension (Phase 3 replication). OR=Odds Ratio, 95%CI=95% Confidence Intervals, *P*=*P*-value.

SNP	CHR	Position	r^2	REF/ ALT Allele	EUR freq	Location	Refseq genes	GERP	Phast Cons	Si-Phy	Proteins bound	Promoter histone marks	Enhancer histone marks	DNase	eQTL tissues	Motifs changed
rs9306894	2	20878105	0.97	A/G	0.36	intergenic	GDF7(dist=6855), C2orf43(dist=6713)						GM12878	NHDF-neo		5 altered motifs
rs9306895	2	20878153	0.97	T/C	0.36	intergenic	GDF7(dist=6903), C2orf43(dist=6665)						GM12878	NHDF-neo		HMG-IY, Lhx3, Pou3f3
rs3072	**2**	**20878406**	**1**	**T/C**	**0.36**	**intergenic**	**GDF7(dist=7156), C2orf43(dist=6412)**						**GM12878**			**GATA, Gfi1**
rs7255	2	20878820	0.60	T/C	0.53	intergenic	GDF7(dist=7570), C2orf43(dist=5998)	2.28	Score 501; lod 145				GM12878			GZF1, Gm397, PLZF
rs10193919	2	20880833	0.92	C/T	0.35	intergenic	GDF7(dist=9583), C2orf43(dist=3985)									6 altered motifs
rs2289081	2	20881840	0.87	G/C	0.36	intergenic	GDF7(dist=10590), C2orf43(dist=2978)									4 altered motifs
rs13394027	2	20882056	0.50	G/A	0.23	intergenic	GDF7(dist=10806), C2orf43(dist=2762)				CTCF, RAD21, SMC3			30 cell types	4 eQTL tissues	Pbx3
rs10170771	2	20883216	0.47	T/C	0.25	intergenic	GDF7(dist=11966), C2orf43(dist=1602)								3 eQTL tissues	4 altered motifs
rs12622106	2	20883561	0.47	C/T	0.25	intergenic	GDF7(dist=12311), C2orf43(dist=1257)				CJUN, JUND				3 eQTL tissues	
rs10171934	2	20884546	0.48	A/C	0.25	downstream	C2orf43				FOXA1, P300		HepG2	HepG2		CEBPA

rs10182643	2	20884586	0.48	G/C	0.25	downstream	C2orf43		FOXA1, P300, SP1	HepG2	HepG2	3 eQTL tissues	NRSF, STAT
rs13385191	2	20888265	0.46	A/G	0.25	intronic	C2orf43	2.86				3 eQTL tissues	Esr2
rs1437405	2	20929067	0.42	T/C	0.32	intronic	C2orf43			HepG2, Huvec		Gibbs Frontal Cortex	22 altered motifs
rs2046325	2	20939706	0.41	T/C	0.32	intronic	C2orf43						12 altered motifs

Table 11.5: Functional annotation of SNPs in LD (r^2>0.4) with rs3072 using data from HaploReg and ANNOVAR. Location, distance from Refseq genes and GERP, PhastCons and SiPhy scores were obtained through annovar. The following files were downloaded from annovar: hg19_refGene.txt, hg19_gerp++gt2.txt, hg19_phastConsElements46way.txt and ljb2_siphy.txt. All SNPs with blank GERP scores map to a location that score <2, therefore is regarded as not being evolutionarily conserved. All SNPs with blank Phastcons scores represent SNPs that do not map to conserved regions. None of the SNPs scored according to SiPhy database (all blank). The Human Protein Atlas showed C2orf43 to be expressed at moderate levels in normal squamous oesophageal epithelial cells and normal glandular stomach, but there was no expression data for the secreted protein GDF7.

SNP	CHR	Position	r^2	REF/ ALT Allele	EUR freq	Location	Refseq genes	GERP	Phast Cons	Si-Phy	Proteins bound	Promoter histone marks	Enhancer histone marks	DNAse	eQTL tissues	Motifs changed
rs1920562	12	114660658	0.62	T/C	0.35	intergenic	RBM19(dist=256482), TBX5(dist=131077)		Score 556; lod 243				H1, NHLF			5 altered motifs
rs12828548	12	114661066	0.62	C/G	0.35	intergenic	RBM19(dist=256890), TBX5(dist=130669)	2.1					H1			Foxj1, Rad21, Zfp105
rs11066998	12	114661789	0.6	A/T	0.35	intergenic	RBM19(dist=257613), TBX5(dist=129946)									Bach2
rs2555009	12	114666099	0.51	A/G	0.52	intergenic	RBM19(dist=261923), TBX5(dist=125636)						HepG2			
rs11067002	12	114667046	0.83	T/A	0.34	intergenic	RBM19(dist=262870), TBX5(dist=124689)									CIZ, Evi-1
rs2555019	12	114668618	0.52	T/C	0.52	intergenic	RBM19(dist=264442), TBX5(dist=123117)	2.13								4 altered motifs
rs1920568	12	114669732	0.58	G/C	0.49	intergenic	RBM19(dist=265556), TBX5(dist=122003)	2.29								EWSR1F LI1, Ets, STAT
rs2555016	12	114670663	0.52	T/G	0.52	intergenic	RBM19(dist=266487), TBX5(dist=121072)									BRCA1, Evi-1, PEBP
rs2701109	12	114671825	0.52	T/A	0.52	intergenic	RBM19(dist=267649), TBX5(dist=119910)									5 altered motifs
rs10850292	12	114672508	0.47	C/A	0.23	intergenic	RBM19(dist=268332), TBX5(dist=119227)									7 altered motifs
rs1247943	12	114673421	0.53	G/A	0.52	intergenic	RBM19(dist=269245), TBX5(dist=118314)									RFX5

rs1247942	12	114673723	0.96	G/C	0.37	intergenic	RBM19(dist=269547), TBX5(dist=118012)					8 altered motifs
rs2555015	12	114673774	0.64	T/C	0.47	intergenic	RBM19(dist=269598), TBX5(dist=117961)					Nkx2, Pbx3
rs2701108	**12**	**114674261**	**1**	**T/C**	**0.36**	**intergenic**	**RBM19(dist=270085), TBX5(dist=117474)**					**4 altered motifs**
rs1270886	12	114676470	0.54	C/T	0.57	intergenic	RBM19 (dist=272294), TBX5(dist=115265)			K562		NRSF
rs1265496	12	114676983	0.55	C/T	0.43	intergenic	RBM19(dist=272807), TBX5(dist=114752)					Pax-4
rs2555014	12	114677491	0.55	G/T	0.43	intergenic	RBM19(dist=273315), TBX5(dist=114244)					Hoxb6, NRSF, Pdx1
rs2555013	12	114678318	0.56	T/C	0.5	intergenic	RBM19(dist=274142), TBX5(dist=113417)					4 altered motifs
rs7980132	12	114678673	0.41	A/G	0.19	intergenic	RBM19(dist=274497), TBX5(dist=113062)					
rs2555012	12	114678725	0.52	C/T	0.44	intergenic	RBM19(dist=274549), TBX5(dist=113010)				Osteobl	GR,RXRA
rs2252414	12	114679137	0.52	G/A	0.44	intergenic	RBM19(dist=274961), TBX5(dist=112598)					12 altered motifs
rs1950090	12	114680189	0.42	A/G	0.58	intergenic	RBM19(dist=276013), TBX5(dist=111546)	BCL3, USF1		K562	HFF-Myc, NT2-D1	10 altered motifs
rs11067013	12	114681027	0.41	C/T	0.19	intergenic	RBM19(dist=276851), TBX5(dist=110708)					4 altered motifs

rs1270885	12	114681552	0.51	A/G	0.44	intergenic	RBM19(dist=277376), TBX5(dist=110183)				CDP,Irx, RXRA
rs1247940	12	114682651	0.52	T/C	0.44	intergenic	RBM19(dist=278475), TBX5(dist=109084)				Pax-5
rs34388546	12	114683214	0.41	C/T	0.19	intergenic	RBM19(dist=279038), TBX5(dist=108521)		H1		Irx
rs2252923	12	114683320	0.48	A/G	0.42	intergenic	RBM19(dist=279144), TBX5(dist=108415)		H1		HNF4, Pax-2, RXRA
rs2252924	12	114683323	0.51	C/A	0.43	intergenic	RBM19(dist=279147), TBX5(dist=108412)		H1		Pax-2
rs1247938	12	114683568	0.52	G/A	0.44	intergenic	RBM19(dist=279392), TBX5(dist=108167)	CTCF, RAD21		12 cell types	NRSF, Spz1
rs1269789	12	114684542	0.52	T/C	0.44	intergenic	RBM19(dist=280366), TBX5(dist=107193)				Irx
rs2253207	12	114685437	0.51	T/C	0.44	intergenic	RBM19(dist=281261), TBX5(dist=106298)				Nkx2
rs1270884	12	114685571	0.54	A/G	0.5	intergenic	RBM19(dist=281395), TBX5(dist=106164)				SRF, TFIIA
rs4007267	12	114685668	0.45	A/T	NA	intergenic	RBM19(dist=281492), TBX5(dist=106067)				21 altered motifs
rs2555004	12	114686645	0.58	G/A	0.49	intergenic	RBM19(dist=282469), TBX5(dist=105090)	NRSF			4 altered motifs
rs1247928	12	114686840	0.52	T/C	0.44	intergenic	RBM19(dist=282664), TBX5(dist=104895)				YY1, Zfp691
rs1247927	12	114687056	0.52	T/C	0.44	intergenic	RBM19(dist=282880), TBX5(dist=104679)				5 altered motifs

rs1247926	12	114687311	0.52	T/C	0.44	intergenic	RBM19(dist=283135), TBX5(dist=104424)		Foxj1, GR, STAT
rs2701111	12	114664920	0.42	G/A	0.46	intergenic	RBM19(dist=260744), TBX5(dist=126815)		5 altered motifs
rs11067004	12	114671418	0.47	A/G	0.23	intergenic	RBM19(dist=267242), TBX5(dist=120317)		4 altered motifs
rs201604604	12	114676112	0.53	T/ TAC	0.43	intergenic	RBM19(dist=271936), TBX5(dist=115623)	K562	4 altered motifs
rs11382177	12	114676113	0.52	A/AC	0.44	intergenic	RBM19(dist=271937), TBX5(dist=115622)	K562	5 altered motifs
rs201882298	12	114683319	0.51	T/TG	0.43	intergenic	RBM19(dist=279143), TBX5(dist=108416)	H1	HNF4, Pax-2
rs200186585	12	114683322	0.49	AC/A	0.42	intergenic	RBM19(dist=279146), TBX5(dist=108413)	H1	Pax-2
rs10706438	12	114685657	0.44	AT/A	0.41	intergenic	RBM19(dist=281481), TBX5(dist=106078)		12 altered motifs
rs201288186	12	114685665	0.42	TTA/ T	0.41	intergenic	RBM19(dist=281489), TBX5(dist=106070)		20 altered motifs
rs35484355	12	114686244	0.51	GC/ G	0.43	intergenic	RBM19(dist=282068), TBX5(dist=105491)		LBP-1

Table 11.6: Functional annotation of SNPs in LD (r^2>0.4) with rs2701108 using data from HaploReg and ANNOVAR. Location, distance from Refseq genes and GERP, PhastCons and SiPhy scores were obtained through ANNOVAR. The following files were downloaded from ANNOVAR: hg19_refGene.txt, hg19_gerp++gt2.txt, hg19_phastConsElements46way.txt and ljb2_siphy.txt. All SNPs with blank GERP scores map to a location that score <2, therefore is regarded as not being evolutionarily conserved. All SNPs with blank PhastCons scores represent SNPs that do not map to conserved regions. None of the SNPs scored according to SiPhy database (all blank). The Human Protein Atlas showed TBX5 and RBM19 to be expressed at moderate levels in normal squamous oesophageal epithelial cells and normal glandular stomach cells.

12. LIST OF ABBREVIATIONS

58C	1958 Birth Cohort
95%CI	95% Confidence Intervals
APC	Adenomatous Polyposis Coli
AspECT	Aspirin Esomeprazole Chemoprevention Trial
BE	Barrett's Oesophagus
BEACON/BEAGESS	Barrett's and Esophageal Adenocarcinoma Consortium
BMI	Body Mass Index
CEU population	Utah residents with Northern and Western European ancestry
CFR	Colon Cancer Family Registry
ChOPIN	Chemoprevention Of Premalignant Intestinal Neoplasia
Chr	Chromosome
CI	Confidence Interval
CNV	Copy Number Variation
CoRGI	Colorectal Tumour Gene Identification
dbSNP	The Single Nucleotide Polymorphism Database
DHH	Desert Hedgehog
dNDP	Deoxyribonucleotide Diphosphate
dNTP	Deoxyribonucleotide Triphosphate
EA	Effect Allele
EAC	Oesophageal Adenocarcinoma
EAGLE	Esophageal Adenocarcinoma GenEtics Consortium
EMT	Epithelial to Mesenchymal Transition
ENCODE	Encyclopedia of DNA Elements
eQTL	Expression Quantitative Trait Loci
FHCRC	Fred Hutchinson Cancer Research Centre

FRET	Fluoresence Resonance Energy Transfer
gDNA	Genomic DNA
GEJ	Gastro-Oesophageal Junction
GEO	Gene Expression Omnibus
GERD	Gastro-Oesophageal Reflux Disease
GI	Gasto-Intestinal
GLACIER	Genetics of LobulAr Carcinoma In situ in EuRope
GWAS	Genome Wide Association Analysis
H.pylori	Helicobacter pylori
HANDEL	Histological Assessment of Neoplasia; Diagnosis and AnaLyses
hg19	Human Genome build 19
HGD	High Grade Dysplasia
HH	Hedgehog
HOS	Holt-Oram Syndrome
HOX	Homeobox
IdU	Iododeoxyuridine
IHH	Indian Hedgehog
IM	Intestinal Metaplasia
IPOD	Inherited Predisposition of Oesophageal Diseases
KASPar	Kompetitive Allele Specific PCR genotyping system
LD	Linkage Disequilibrium
LGD	Low Grade Dysplasia
LOH	Loss Of Heterozygosity
LOS	Lower Oesophgeal Spincter
LSBE	Long Segment Barrett's Oesophagus
MAF	Minor Allele Frequency
MALDI-TOF	Matrix Assisted Laser Desorption/Ionization-Time Of Flight
MAPK	Mitogen-Activated Protein Kinase

MCS	Metaplasia-dysplasia-adenocarcinoma sequence
MREC	Multicentre Research Ethics Committee
NCBI	National Center for Biotechnology Information
NCDS	The National Child Development Study
NEA	Non-Effect Allele
NHS	National Health Service
NSAID	Non-Steroidal Anti-Inflammatory Drug
OR	Odds Ratio
PC	Principal Component
PCA	Principal component analysis
PCR	Polymerase Chain Reaction
PoBI	People of the British Isles
PPI	Proton Pump Inhibitors
QC	Quality Control
RA	Retinoic Acid
SAP	Shrimp Alkaline Phospatase
SBE	Single Base Extension
SE	Standard Error
SHH	Sonic Hedgehog
SNAP	SNP Annotation and Proxy Search
SNP	Single Nucleotide Polymorphism
SSBE	Short Segment Barrett's Oesophagus
TGFβ	Transforming Growth Factor β
TOF	Tetralogy of Fallot
UKBS	National Blood Service
UKREP1	UK Replication 1
UKREP2	UK Replication 2
UKREP3	UK Replication 3

UOS	Upper Oesophageal Sphincter
WHR	Waist-Hip Ratio
WTCCC2	Wellcome Trust Case Control Consortium 2
WTCHG	Wellcome Trust Centre for Human Genetics
WTSI	Wellcome Trust Sanger Institute

13. PUBLICATIONS

Palles, C., et al., Polymorphisms near TBX5 and GDF7 are associated with increased risk for Barrett's esophagus. Gastroenterology, 2015. 148(2): p. 367-78.

14. URLs

1000 Genomes: http://www.1000genomes.org/

ANNOVAR: http://www.openbioinformatics.org/annovar/

GTOOL: http://www.well.ox.ac.uk/~cfreeman/software/gwas/gtool.html

GWAMA: http://www.well.ox.ac.uk/gwama/index.shtml

HAPLOREG: http://www.broadinstitute.org/mammals/haploreg/haploreg.php

Human Protein Atlas: http://www.proteinatlas.org

IMPUTE2: http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

LOCUS ZOOM: http://csg.sph.umich.edu/locuszoom/

PLINK: http://pngu.mgh.harvard.edu/~purcell/plink/

Regulome DB: http://www.regulomedb.org/

SCAN: http://www.scandb.org/newinterface/about.html

SNPTEST: https://mathgen.stats.ox.ac.uk/genetics_software/snptest/old/snptest.html

UCSC Genome Browser: https://genome.ucsc.edu

15. REFERENCES

1. Palles, C., et al., *Polymorphisms near TBX5 and GDF7 are associated with increased risk for Barrett's esophagus.* Gastroenterology, 2015. **148**(2): p. 367-78.
2. Su, Z., et al., *Common variants at the MHC locus and at chromosome 16q24.1 predispose to Barrett's esophagus.* Nat Genet, 2012. **44**(10): p. 1131-6.
3. Levine, D.M., et al., *A genome-wide association study identifies new susceptibility loci for esophageal adenocarcinoma and Barrett's esophagus.* Nat Genet, 2013. **45**(12): p. 1487-93.
4. Gabriel, S., L. Ziaugra, and D. Tabbaa, *SNP genotyping using the Sequenom MassARRAY iPLEX platform.* Curr Protoc Hum Genet, 2009. **Chapter 2**: p. Unit 2.12.
5. Kent, W.J., et al., *The human genome browser at UCSC.* Genome Res, 2002. **12**(6): p. 996-1006.
6. Phillips, W.A., et al., *Barrett's esophagus.* J Gastroenterol Hepatol, 2011. **26**(4): p. 639-48.
7. Jankowski, J., et al., *Diagnosis and management of Barrett's oesophagus.* BMJ, 2010. **341**: p. c4551.
8. BARRETT, N.R., *Chronic peptic ulcer of the oesophagus and 'oesophagitis'.* Br J Surg, 1950. **38**(150): p. 175-82.
9. Theisen, J., et al., *Chronology of the Barrett's metaplasia-dysplasia-carcinoma sequence.* Dis Esophagus, 2004. **17**(1): p. 67-70.
10. Gilbert, E.W., et al., *Barrett's esophagus: a review of the literature.* J Gastrointest Surg, 2011. **15**(5): p. 708-18.
11. Jankowski, J.A., et al., *Molecular evolution of the metaplasia-dysplasia-adenocarcinoma sequence in the esophagus.* Am J Pathol, 1999. **154**(4): p. 965-73.
12. Deviere, J., *Barrett's oesophagus: the new endoscopic modalities have a future.* Gut, 2005. **54 Suppl 1**: p. i33-7.
13. Corley, D.A. and P.A. Buffler, *Oesophageal and gastric cardia adenocarcinomas: analysis of regional variation using the Cancer Incidence in Five Continents database.* Int J Epidemiol, 2001. **30**(6): p. 1415-25.
14. Gerson, L.B., K. Shetler, and G. Triadafilopoulos, *Prevalence of Barrett's esophagus in asymptomatic individuals.* Gastroenterology, 2002. **123**(2): p. 461-7.
15. Hayeck, T.J., et al., *The prevalence of Barrett's esophagus in the US: estimates from a simulation model confirmed by SEER data.* Dis Esophagus, 2010. **23**(6): p. 451-7.
16. Ronkainen, J., et al., *Prevalence of Barrett's esophagus in the general population: an endoscopic study.* Gastroenterology, 2005. **129**(6): p. 1825-31.
17. Zagari, R.M., et al., *Gastro-oesophageal reflux symptoms, oesophagitis and Barrett's oesophagus in the general population: the Loiano-Monghidoro study.* Gut, 2008. **57**(10): p. 1354-9.
18. Evans, J.A., et al., *The role of endoscopy in Barrett's esophagus and other premalignant conditions of the esophagus.* Gastrointest Endosc, 2012. **76**(6): p. 1087-94.
19. Garud, S.S., et al., *Diagnosis and management of Barrett's esophagus for the endoscopist.* Therap Adv Gastroenterol, 2010. **3**(4): p. 227-38.
20. Jankowski, J.A., et al., *Barrett's metaplasia.* Lancet, 2000. **356**(9247): p. 2079-85.
21. Nilsson, M., et al., *Body mass and reflux oesophagitis: an oestrogen-dependent association?* Scand J Gastroenterol, 2002. **37**(6): p. 626-30.

22. Harnish, D.C., *Estrogen receptor ligands in the control of pathogenic inflammation.* Curr Opin Investig Drugs, 2006. **7**(11): p. 997-1001.
23. Shaheen, N.J. and J.E. Richter, *Barrett's oesophagus.* Lancet, 2009. **373**(9666): p. 850-61.
24. Kubo, A., et al., *Cigarette smoking and the risk of Barrett's esophagus.* Cancer Causes Control, 2009. **20**(3): p. 303-11.
25. Balasubramanian, G., et al., *Cigarette smoking is a modifiable risk factor for Barrett's oesophagus.* United European Gastroenterol J, 2013. **1**(6): p. 430-7.
26. Gatenby, P. and Y. Soon, *Barrett's oesophagus: Evidence from the current meta-analyses.* World J Gastrointest Pathophysiol, 2014. **5**(3): p. 178-87.
27. Kamat, P., et al., *Exploring the association between elevated body mass index and Barrett's esophagus: a systematic review and meta-analysis.* Ann Thorac Surg, 2009. **87**(2): p. 655-62.
28. Kubo, A., et al., *Sex-specific associations between body mass index, waist circumference and the risk of Barrett's oesophagus: a pooled analysis from the international BEACON consortium.* Gut, 2013. **62**(12): p. 1684-91.
29. Kramer, J.R., et al., *Waist-to-hip ratio, but not body mass index, is associated with an increased risk of Barrett's esophagus in white men.* Clin Gastroenterol Hepatol, 2013. **11**(4): p. 373-381.e1.
30. Edelstein, Z.R., et al., *Risk factors for Barrett's esophagus among patients with gastroesophageal reflux disease: a community clinic-based case-control study.* Am J Gastroenterol, 2009. **104**(4): p. 834-42.
31. DeVault, K.R. and D.O. Castell, *Updated guidelines for the diagnosis and treatment of gastroesophageal reflux disease. The Practice Parameters Committee of the American College of Gastroenterology.* Am J Gastroenterol, 1999. **94**(6): p. 1434-42.
32. Vakil, N., et al., *The Montreal definition and classification of gastroesophageal reflux disease: a global evidence-based consensus.* Am J Gastroenterol, 2006. **101**(8): p. 1900-20; quiz 1943.
33. Sontag, S.J., *Defining GERD.* Yale J Biol Med, 1999. **72**(2-3): p. 69-80.
34. ZOLLINGER, R.M. and E.H. ELLISON, *Primary peptic ulcerations of the jejunum associated with islet cell tumors of the pancreas.* Ann Surg, 1955. **142**(4): p. 709-23; discussion, 724-8.
35. Lamers, C.B. and J.H. Van Tongeren, *Serum gastrin response to acute and chronic hypercalcaemia in man: studies on the value of calcium stimulated serum gastrin levels in the diagnosis of Zollinger-Ellison syndrome.* Eur J Clin Invest, 1977. **7**(4): p. 315-7.
36. Lee, P., J. Bruni, and S. Sukenik, *Neurological manifestations in systemic sclerosis (scleroderma).* J Rheumatol, 1984. **11**(4): p. 480-3.
37. Katz, P.O., L.B. Gerson, and M.F. Vela, *Guidelines for the diagnosis and management of gastroesophageal reflux disease.* Am J Gastroenterol, 2013. **108**(3): p. 308-28; quiz 329.
38. Chiba, N., *Proton pump inhibitors in acute healing and maintenance of erosive or worse esophagitis: a systematic overview.* Can J Gastroenterol, 1997. **11 Suppl B**: p. 66B-73B.
39. Törüner, M., et al., *The effect of rabeprazole alone or in combination with H2 receptor blocker on intragastric pH: a pilot study.* Turk J Gastroenterol, 2004. **15**(4): p. 225-8.
40. Kala, Z., et al., *Polymorphisms of glutathione S-transferase M1, T1 and P1 in patients with reflux esophagitis and Barrett's esophagus.* J Hum Genet, 2007. **52**(6): p. 527-34.
41. Kubo, A., et al., *Dietary antioxidants, fruits, and vegetables and the risk of Barrett's esophagus.* Am J Gastroenterol, 2008. **103**(7): p. 1614-23; quiz 1624.
42. Kubo, A., et al., *Effects of dietary fiber, fats, and meat intakes on the risk of Barrett's esophagus.* Nutr Cancer, 2009. **61**(5): p. 607-16.

43. MacDonald, K., et al., *A polymorphic variant of the insulin-like growth factor type I receptor gene modifies risk of obesity for esophageal adenocarcinoma.* Cancer Epidemiol, 2009. **33**(1): p. 37-40.
44. Menke, V., et al., *Myo9B is associated with an increased risk of Barrett's esophagus and esophageal adenocarcinoma.* Scand J Gastroenterol, 2012. **47**(12): p. 1422-8.
45. Menke, V., et al., *NcoI TNF-β gene polymorphism and TNF expression are associated with an increased risk of developing Barrett's esophagus and esophageal adenocarcinoma.* Scand J Gastroenterol, 2012. **47**(4): p. 378-86.
46. Orloff, M., et al., *Germline mutations in MSR1, ASCC1, and CTHRC1 in patients with Barrett esophagus and esophageal adenocarcinoma.* JAMA, 2011. **306**(4): p. 410-9.
47. Ren, D., et al., *Single nucleotide polymorphisms of caudal type homeobox 1 and 2 are associated with Barrett's esophagus.* Dig Dis Sci, 2014. **59**(1): p. 57-63.
48. Fischbach, L.A., et al., *Association between Helicobacter pylori and Barrett's esophagus: a case-control study.* Am J Gastroenterol, 2014. **109**(3): p. 357-68.
49. Corley, D.A., et al., *Helicobacter pylori infection and the risk of Barrett's oesophagus: a community-based study.* Gut, 2008. **57**(6): p. 727-33.
50. Anderson, L.A., et al., *Relationship between Helicobacter pylori infection and gastric atrophy and the stages of the oesophageal inflammation, metaplasia, adenocarcinoma sequence: results from the FINBAR case-control study.* Gut, 2008. **57**(6): p. 734-9.
51. Anderson, L.A., et al., *Nonsteroidal anti-inflammatory drugs and the esophageal inflammation-metaplasia-adenocarcinoma sequence.* Cancer Res, 2006. **66**(9): p. 4975-82.
52. Wang, F., Z.S. Lv, and Y.K. Fu, *Nonsteroidal anti-inflammatory drugs and esophageal inflammation - Barrett's esophagus - adenocarcinoma sequence: a meta-analysis.* Dis Esophagus, 2010.
53. Fitzgerald, R.C., et al., *British Society of Gastroenterology guidelines on the diagnosis and management of Barrett's oesophagus.* Gut, 2014. **63**(1): p. 7-42.
54. Seewald, S., T.L. Ang, and N. Soehendra, *Endoscopic mucosal resection of Barrett's oesophagus containing dysplasia or intramucosal cancer.* Postgrad Med J, 2007. **83**(980): p. 367-72.
55. Bright, T., et al., *Randomized trial of argon plasma coagulation versus endoscopic surveillance for barrett esophagus after antireflux surgery: late results.* Ann Surg, 2007. **246**(6): p. 1016-20.
56. Barr, H., et al., *Eradication of high-grade dysplasia in columnar-lined (Barrett's) oesophagus by photodynamic therapy with endogenously generated protoporphyrin IX.* Lancet, 1996. **348**(9027): p. 584-5.
57. Wang, K.K., R.E. Sampliner, and P.P.C.o.t.A.C.o. Gastroenterology, *Updated guidelines 2008 for the diagnosis, surveillance and therapy of Barrett's esophagus.* Am J Gastroenterol, 2008. **103**(3): p. 788-97.
58. Prach, A.T., et al., *Increasing incidence of Barrett's oesophagus: education, enthusiasm, or epidemiology?* Lancet, 1997. **350**(9082): p. 933.
59. Moayyedi, P., et al., *Mortality rates in patients with Barrett's oesophagus.* Aliment Pharmacol Ther, 2008. **27**(4): p. 316-20.
60. Gurski, R.R., et al., *Barrett's esophagus can and does regress after antireflux surgery: a study of prevalence and predictive features.* J Am Coll Surg, 2003. **196**(5): p. 706-12; discussion 712-3.
61. Peters, F.T., et al., *Endoscopic regression of Barrett's oesophagus during omeprazole treatment; a randomised double blind study.* Gut, 1999. **45**(4): p. 489-94.
62. Fitzgerald, R.C., *Molecular basis of Barrett's oesophagus and oesophageal adenocarcinoma.* Gut, 2006. **55**(12): p. 1810-20.

63. Hamilton, S.R. and J.H. Yardley, *Regnerative of cardiac type mucosa and acquisition of Barrett mucosa after esophagogastrostomy.* Gastroenterology, 1977. **72**(4 Pt 1): p. 669-75.
64. Li, H., et al., *Mechanisms of columnar metaplasia and squamous regeneration in experimental Barrett's esophagus.* Surgery, 1994. **115**(2): p. 176-81.
65. Koak, Y. and M. Winslet, *Changing role of in vivo models in columnar-lined lower esophagus.* Dis Esophagus, 2002. **15**(4): p. 271-7.
66. Gillen, P., et al., *Experimental columnar metaplasia in the canine oesophagus.* Br J Surg, 1988. **75**(2): p. 113-5.
67. Croagh, D., et al., *Identification of candidate murine esophageal stem cells using a combination of cell kinetic studies and cell surface markers.* Stem Cells, 2007. **25**(2): p. 313-8.
68. Seery, J.P., *Stem cells of the oesophageal epithelium.* J Cell Sci, 2002. **115**(Pt 9): p. 1783-9.
69. Chang, C.L., et al., *Retinoic acid-induced glandular differentiation of the oesophagus.* Gut, 2007. **56**(7): p. 906-17.
70. Coad, R.A., et al., *On the histogenesis of Barrett's oesophagus and its associated squamous islands: a three-dimensional study of their morphological relationship with native oesophageal gland ducts.* J Pathol, 2005. **206**(4): p. 388-94.
71. Leedham, S.J., et al., *Individual crypt genetic heterogeneity and the origin of metaplastic glandular epithelium in human Barrett's oesophagus.* Gut, 2008. **57**(8): p. 1041-8.
72. Tosh, D. and J.M. Slack, *How cells change their phenotype.* Nat Rev Mol Cell Biol, 2002. **3**(3): p. 187-94.
73. JOHNS, B.A., *Developmental changes in the oesophageal epithelium in man.* J Anat, 1952. **86**(4): p. 431-42.
74. Yu, W.Y., J.M. Slack, and D. Tosh, *Conversion of columnar to stratified squamous epithelium in the developing mouse oesophagus.* Dev Biol, 2005. **284**(1): p. 157-70.
75. Bajpai, M., et al., *Repeated exposure to acid and bile selectively induces colonic phenotype expression in a heterogeneous Barrett's epithelial cell line.* Lab Invest, 2008. **88**(6): p. 643-51.
76. Kosinski, C., et al., *Indian hedgehog regulates intestinal stem cell fate through epithelial-mesenchymal interactions during development.* Gastroenterology, 2010. **139**(3): p. 893-903.
77. Wang, X., et al., *Residual embryonic cells as precursors of a Barrett's-like metaplasia.* Cell, 2011. **145**(7): p. 1023-35.
78. Quante, M., et al., *Bile acid and inflammation activate gastric cardia stem cells in a mouse model of Barrett-like metaplasia.* Cancer Cell, 2012. **21**(1): p. 36-51.
79. Lavery, D.L., et al., *The stem cell organisation, and the proliferative and gene expression profile of Barrett's epithelium, replicates pyloric-type gastric glands.* Gut, 2014.
80. Katoh, M., *Networking of WNT, FGF, Notch, BMP, and Hedgehog signaling pathways during carcinogenesis.* Stem Cell Rev, 2007. **3**(1): p. 30-8.
81. Litingtung, Y., et al., *Sonic hedgehog is essential to foregut development.* Nat Genet, 1998. **20**(1): p. 58-61.
82. Madison, B.B., et al., *Epithelial hedgehog signals pattern the intestinal crypt-villus axis.* Development, 2005. **132**(2): p. 279-89.
83. Rubin, L.L. and F.J. de Sauvage, *Targeting the Hedgehog pathway in cancer.* Nat Rev Drug Discov, 2006. **5**(12): p. 1026-33.
84. Wang, D.H., et al., *Aberrant epithelial-mesenchymal Hedgehog signaling characterizes Barrett's metaplasia.* Gastroenterology, 2010. **138**(5): p. 1810-22.
85. Masuda, S., *Dysfunctional transforming growth factor-β signaling with constitutively active notch signaling in Barrett's esophageal adenocarcinoma.* Cancer, 2012. **118**(7): p. 1956-7; author reply 1957-8.

86. Mendelson, J., et al., *Dysfunctional transforming growth factor-β signaling with constitutively active Notch signaling in Barrett's esophageal adenocarcinoma.* Cancer, 2011. **117**(16): p. 3691-702.
87. Souza, R.F., et al., *Acid exposure activates the mitogen-activated protein kinase pathways in Barrett's esophagus.* Gastroenterology, 2002. **122**(2): p. 299-307.
88. Jaiswal, K., et al., *Bile salt exposure increases proliferation through p38 and ERK MAPK pathways in a non-neoplastic Barrett's cell line.* Am J Physiol Gastrointest Liver Physiol, 2006. **290**(2): p. G335-42.
89. Brittan, M. and N.A. Wright, *Gastrointestinal stem cells.* J Pathol, 2002. **197**(4): p. 492-509.
90. Li, X., et al., *Single nucleotide polymorphism-based genome-wide chromosome copy change, loss of heterozygosity, and aneuploidy in Barrett's esophagus neoplastic progression.* Cancer Prev Res (Phila), 2008. **1**(6): p. 413-23.
91. Paulson, T.G., et al., *Chromosomal instability and copy number alterations in Barrett's esophagus and esophageal adenocarcinoma.* Clin Cancer Res, 2009. **15**(10): p. 3305-14.
92. Chaves, P., et al., *Chromosomal analysis of Barrett's cells: demonstration of instability and detection of the metaplastic lineage involved.* Mod Pathol, 2007. **20**(7): p. 788-96.
93. Lohnes, D., *The Cdx1 homeodomain protein: an integrator of posterior signaling in the mouse.* Bioessays, 2003. **25**(10): p. 971-80.
94. Bonner, C.A., S.K. Loftus, and J.J. Wasmuth, *Isolation, characterization, and precise physical localization of human CDX1, a caudal-type homeobox gene.* Genomics, 1995. **28**(2): p. 206-11.
95. Meyer, B.I. and P. Gruss, *Mouse Cdx-1 expression during gastrulation.* Development, 1993. **117**(1): p. 191-203.
96. Freund, J.N., et al., *The Cdx-1 and Cdx-2 homeobox genes in the intestine.* Biochem Cell Biol, 1998. **76**(6): p. 957-69.
97. Moons, L.M., et al., *The homeodomain protein CDX2 is an early marker of Barrett's oesophagus.* J Clin Pathol, 2004. **57**(10): p. 1063-8.
98. Casson, A.G., et al., *p53 gene mutations in Barrett's epithelium and esophageal cancer.* Cancer Res, 1991. **51**(16): p. 4495-9.
99. Fléjou, J.F., et al., *Overexpression of the p53 tumor suppressor gene product in esophageal and gastric carcinomas.* Pathol Res Pract, 1994. **190**(12): p. 1141-8.
100. Hall, P.A., et al., *Expression of the p53 homologue p63alpha and DeltaNp63alpha in the neoplastic sequence of Barrett's oesophagus: correlation with morphology and p53 protein.* Gut, 2001. **49**(5): p. 618-23.
101. Daniely, Y., et al., *Critical role of p63 in the development of a normal esophageal and tracheobronchial epithelium.* Am J Physiol Cell Physiol, 2004. **287**(1): p. C171-81.
102. Bian, Y.S., et al., *p16 inactivation by methylation of the CDKN2A promoter occurs early during neoplastic progression in Barrett's esophagus.* Gastroenterology, 2002. **122**(4): p. 1113-21.
103. Werner, M., et al., *The molecular pathology of Barrett's esophagus.* Histol Histopathol, 1999. **14**(2): p. 553-9.
104. Wu, T.T., et al., *Genetic alterations in Barrett esophagus and adenocarcinomas of the esophagus and esophagogastric junction region.* Am J Pathol, 1998. **153**(1): p. 287-94.
105. Howard, S., et al., *A positive role of cadherin in Wnt/β-catenin signalling during epithelial-mesenchymal transition.* PLoS One, 2011. **6**(8): p. e23899.
106. Izakovicova Holla, L., et al., *Haplotypes of the IL-1 gene cluster are associated with gastroesophageal reflux disease and Barrett's esophagus.* Hum Immunol, 2013. **74**(9): p. 1161-9.

107. Menke, V., et al., *Functional single-nucleotide polymorphism of epidermal growth factor is associated with the development of Barrett's esophagus and esophageal adenocarcinoma.* J Hum Genet, 2012. **57**(1): p. 26-32.
108. Babar, M., et al., *Genes of the interleukin-18 pathway are associated with susceptibility to Barrett's esophagus and esophageal adenocarcinoma.* Am J Gastroenterol, 2012. **107**(9): p. 1331-41.
109. van de Winkel, A., et al., *Expression, localization and polymorphisms of the nuclear receptor PXR in Barrett's esophagus and esophageal adenocarcinoma.* BMC Gastroenterol, 2011. **11**: p. 108.
110. McElholm, A.R., et al., *A population-based study of IGF axis polymorphisms and the esophageal inflammation, metaplasia, adenocarcinoma sequence.* Gastroenterology, 2010. **139**(1): p. 204-12.e3.
111. Moons, L.M., et al., *A pro-inflammatory genotype predisposes to Barrett's esophagus.* Carcinogenesis, 2008. **29**(5): p. 926-31.
112. di Martino, E., et al., *The NAD(P)H:quinone oxidoreductase I C609T polymorphism modifies the risk of Barrett esophagus and esophageal adenocarcinoma.* Genet Med, 2007. **9**(6): p. 341-7.
113. Casson, A.G., et al., *Cyclin D1 polymorphism (G870A) and risk for esophageal adenocarcinoma.* Cancer, 2005. **104**(4): p. 730-9.
114. Casson, A.G., et al., *Polymorphisms in DNA repair genes in the molecular pathogenesis of esophageal (Barrett) adenocarcinoma.* Carcinogenesis, 2005. **26**(9): p. 1536-41.
115. Gough, M.D., et al., *Prediction of malignant potential in reflux disease: are cytokine polymorphisms important?* Am J Gastroenterol, 2005. **100**(5): p. 1012-8.
116. van Lieshout, E.M., et al., *Polymorphic expression of the glutathione S-transferase P1 gene and its susceptibility to Barrett's esophagus and esophageal carcinoma.* Cancer Res, 1999. **59**(3): p. 586-9.
117. Briant, L., et al., *Multiple sclerosis susceptibility: population and twin study of polymorphisms in the T-cell receptor beta and gamma genes region. French Group on Multiple Sclerosis.* Autoimmunity, 1993. **15**(1): p. 67-73.
118. Cuvelier, C., et al., *Idiopathic inflammatory bowel diseases: immunological hypothesis.* Acta Gastroenterol Belg, 1994. **57**(5-6): p. 292-9.
119. Kim, B.M., et al., *The stomach mesenchymal transcription factor Barx1 specifies gastric epithelial identity through inhibition of transient Wnt signaling.* Dev Cell, 2005. **8**(4): p. 611-22.
120. Miletich, I., G. Buchner, and P.T. Sharpe, *Barx1 and evolutionary changes in feeding.* J Anat, 2005. **207**(5): p. 619-22.
121. Banham, A.H., et al., *The FOXP1 winged helix transcription factor is a novel candidate tumor suppressor gene on chromosome 3p.* Cancer Res, 2001. **61**(24): p. 8820-9.
122. Katoh, M., et al., *Cancer genetics and genomics of human FOX family genes.* Cancer Lett, 2013. **328**(2): p. 198-206.
123. Shu, W., et al., *Characterization of a new subfamily of winged-helix/forkhead (Fox) genes that are expressed in the lung and act as transcriptional repressors.* J Biol Chem, 2001. **276**(29): p. 27488-97.
124. Jankowski, J. and H. Barr, *Improving surveillance for Barrett's oesophagus: AspECT and BOSS trials provide an evidence base.* BMJ, 2006. **332**(7556): p. 1512.
125. Power, C. and J. Elliott, *Cohort profile: 1958 British birth cohort (National Child Development Study).* Int J Epidemiol, 2006. **35**(1): p. 34-41.
126. Consortium, W.T.C.C., *Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.* Nature, 2007. **447**(7145): p. 661-78.
127. Houlston, R.S., et al., *Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33.* Nat Genet, 2010. **42**(11): p. 973-7.

128. Trynka, G., et al., *Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease.* Nat Genet, 2011. **43**(12): p. 1193-201.
129. Sharma, P., et al., *The development and validation of an endoscopic grading system for Barrett's esophagus: the Prague C & M criteria.* Gastroenterology, 2006. **131**(5): p. 1392-9.
130. Petridis, C., et al., *Germline CDH1 mutations in bilateral lobular carcinoma in situ.* Br J Cancer, 2014. **110**(4): p. 1053-7.
131. Tang, K., et al., *Chip-based genotyping by mass spectrometry.* Proc Natl Acad Sci U S A, 1999. **96**(18): p. 10016-20.
132. *Sequenom Inc.; MassARRAY® iPLEX® Gold – SNP Genotyping; #sq177_v2(092010)*
133. *Infinium® HD Assay Super Protocol Guide- ILLUMINA PROPRIETARY Catalog # WG-901-4002 Part # 11322427 Rev. C*.
134. Steemers, F.J. and K.L. Gunderson, *Whole genome genotyping technologies on the BeadArray platform.* Biotechnol J, 2007. **2**(1): p. 41-9.
135. Adler, A.J., G.B. Wiley, and P.M. Gaffney, *Infinium assay for large-scale SNP genotyping applications.* J Vis Exp, 2013(81): p. e50683.
136. *Technology Spotlight: SNP Genotyping; Infinium® Assay Workflow*.
137. Cortes, A. and M.A. Brown, *Promise and pitfalls of the Immunochip.* Arthritis Res Ther, 2011. **13**(1): p. 101.
138. *How does KASP work;* http://www.lgcgroup.com/kasp/. Available from: http://www.lgcgroup.com/kasp/ - .VCqBo0uppuY.
139. *KASP Manual;* http://www.lgcgroup.com/products/kasp-genotyping-chemistry/kasp-technical-resources/. Available from: http://www.lgcgroup.com/products/kasp-genotyping-chemistry/kasp-technical-resources/.
140. *KASP quick start guide;* http://www.lgcgroup.com/products/kasp-genotyping-chemistry/kasp-technical-resources. Available from: http://www.lgcgroup.com/products/kasp-genotyping-chemistry/kasp-technical-resources.
141. He, C., J. Holme, and J. Anthony, *SNP genotyping: the KASP assay.* Methods Mol Biol, 2014. **1145**: p. 75-86.
142. Purcell S, N.B., Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ & Sham PC, *PLINK: a toolset for whole-genome association and population-based linkage analysis. .* American Journal of Human Genetics, 81., 2007.
143. Mägi, R. and A.P. Morris, *GWAMA: software for genome-wide association meta-analysis.* BMC Bioinformatics, 2010. **11**: p. 288.
144. Marchini, J. and B. Howie, *Genotype imputation for genome-wide association studies.* Nat Rev Genet, 2010. **11**(7): p. 499-511.
145. Howie, B.N., P. Donnelly, and J. Marchini, *A flexible and accurate genotype imputation method for the next generation of genome-wide association studies.* PLoS Genet, 2009. **5**(6): p. e1000529.
146. Ward, L.D. and M. Kellis, *HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants.* Nucleic Acids Res, 2012. **40**(Database issue): p. D930-4.
147. Boyle, A.P., et al., *Annotation of functional variation in personal genomes using RegulomeDB.* Genome Res, 2012. **22**(9): p. 1790-7.
148. Gamazon, E.R., et al., *SCAN: SNP and copy number annotation.* Bioinformatics, 2010. **26**(2): p. 259-62.
149. Wang, K., M. Li, and H. Hakonarson, *ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data.* Nucleic Acids Res, 2010. **38**(16): p. e164.
150. Dunlop, M.G., et al., *Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk.* Nat Genet, 2012. **44**(7): p. 770-6.

151. Duggal, P., et al., *Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies.* BMC Genomics, 2008. **9**: p. 516.
152. Pe'er, I., et al., *Estimation of the multiple testing burden for genomewide association studies of nearly all common variants.* Genet Epidemiol, 2008. **32**(4): p. 381-5.
153. Dudbridge, F. and A. Gusnanto, *Estimation of significance thresholds for genomewide association scans.* Genet Epidemiol, 2008. **32**(3): p. 227-34.
154. Risch, N. and K. Merikangas, *The future of genetic studies of complex human diseases.* Science, 1996. **273**(5281): p. 1516-7.
155. Benevolenskaya, E.V., *Histone H3K4 demethylases are essential in development and differentiation.* Biochem Cell Biol, 2007. **85**(4): p. 435-43.
156. Didon, L., et al., *RFX3 modulation of FOXJ1 regulation of cilia genes in the human airway epithelium.* Respir Res, 2013. **14**: p. 70.
157. Katoh, Y. and M. Katoh, *Comparative integromics on BMP/GDF family.* Int J Mol Med, 2006. **17**(5): p. 951-5.
158. Lou, J., et al., *BMP-12 gene transfer augmentation of lacerated tendon repair.* J Orthop Res, 2001. **19**(6): p. 1199-202.
159. Fu, S.C., et al., *The roles of bone morphogenetic protein (BMP) 12 in stimulating the proliferation and matrix production of human patellar tendon fibroblasts.* Life Sci, 2003. **72**(26): p. 2965-74.
160. Long, Q.Z., et al., *Replication and fine mapping for association of the C2orf43, FOXP4, GPRC6A and RFX6 genes with prostate cancer in the Chinese population.* PLoS One, 2012. **7**(5): p. e37866.
161. Takata, R., et al., *Genome-wide association study identifies five new susceptibility loci for prostate cancer in the Japanese population.* Nat Genet, 2010. **42**(9): p. 751-4.
162. Haveri, H., et al., *Transcription factors GATA-4 and GATA-6 in normal and neoplastic human gastrointestinal mucosa.* BMC Gastroenterol, 2008. **8**: p. 9.
163. Maijgren, S., et al., *Involvement of RFX proteins in transcriptional activation from a Ras-responsive enhancer element.* Arch Dermatol Res, 2004. **295**(11): p. 482-9.
164. Milano, F., et al., *Bone morphogenetic protein 4 expressed in esophagitis induces a columnar phenotype in esophageal squamous cells.* Gastroenterology, 2007. **132**(7): p. 2412-21.
165. Castillo, D., et al., *Activation of the BMP4 pathway and early expression of CDX2 characterize non-specialized columnar metaplasia in a human model of Barrett's esophagus.* J Gastrointest Surg, 2012. **16**(2): p. 227-37; discussion 237.
166. Baban, A., et al., *Identification of TBX5 mutations in a series of 94 patients with Tetralogy of Fallot.* Am J Med Genet A, 2014.
167. Li, Q.Y., et al., *Holt-Oram syndrome is caused by mutations in TBX5, a member of the Brachyury (T) gene family.* Nat Genet, 1997. **15**(1): p. 21-9.
168. Arora, R., R.J. Metzger, and V.E. Papaioannou, *Multiple roles and interactions of Tbx4 and Tbx5 in development of the respiratory system.* PLoS Genet, 2012. **8**(8): p. e1002866.
169. Hasson, P., et al., *Tbx4 and tbx5 acting in connective tissue are required for limb muscle and tendon patterning.* Dev Cell, 2010. **18**(1): p. 148-56.
170. Lorenzen, J.A., et al., *Rbm19 is a nucleolar protein expressed in crypt/progenitor cells of the intestinal epithelium.* Gene Expr Patterns, 2005. **6**(1): p. 45-56.
171. Borozdin, W., et al., *Contiguous hemizygous deletion of TBX5, TBX3, and RBM19 resulting in a combined phenotype of Holt-Oram and ulnar-mammary syndromes.* Am J Med Genet A, 2006. **140A**(17): p. 1880-6.
172. Shu, W., et al., *Foxp2 and Foxp1 cooperatively regulate lung and esophagus development.* Development, 2007. **134**(10): p. 1991-2000.

173. Gyamfi, M.A., et al., *The role of retinoid X receptor alpha in regulating alcohol metabolism.* J Pharmacol Exp Ther, 2006. **319**(1): p. 360-8.
174. Harlaar, N., et al., *Methylation of a CpG site near the ALDH1A2 gene is associated with loss of control over drinking and related phenotypes.* Alcohol Clin Exp Res, 2014. **38**(3): p. 713-21.
175. Pavan, M., et al., *ALDH1A2 (RALDH2) genetic variation in human congenital heart disease.* BMC Med Genet, 2009. **10**: p. 113.
176. Kim, H., et al., *The retinoic acid synthesis gene ALDH1a2 is a candidate tumor suppressor in prostate cancer.* Cancer Res, 2005. **65**(18): p. 8118-24.