UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

FACULTADE DE MEDICINA

DEPARTAMENTO DE ANATOMÍA PATOLÓXICA E CIENCIAS FORENSES

GENOMICS AND PHARMACOGENOMICS OF COLORECTAL CANCER

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

CERES FERNÁNDEZ ROZADILLA

Santiago de Compostela, Mayo 2011

ISBN 978-84-9887-769-4 (Edición digital PDF)

El Profesor Doctor Ángel Carracedo Álvarez, Catedrático de Medicina Legal de la Facultad de Medicina de la Universidad de Santiago de Compostela, y la Doctora Clara Ruiz Ponte, adjunto de la Fundación Pública Galega de Medicina Xenómica

CERTIFICAN:

que la presente Tesis Doctoral: **'Genomics and Pharmacogenomics of Colorectal Cancer'** de la Licenciada en Biología Dña. Ceres Fernández Rozadilla ha sido realizada bajo su dirección, considerándola en condiciones para optar al Grado de Doctor y autorizándola para presentar ante el tribunal correspondiente.

Para que así conste, se expide el presente certificado en Santiago de Compostela a 11 de mayo de 2011

Fdo.: Prof Dr. Angel Carracedo

Fdo.: Clara Ruiz Ponte

CFR´s work, as well as the stay at the Wellcome Trust Centre for Human Genetics, was supported by an FPU grant from the Ministerio de Ciencia e Innovación (07-03883).

Funding for this work included Fondo de Investigacion Sanitaria/FEDER (05/2031, 08/0024, 08/1276, PS09/02368), Xunta de Galicia (PGIDIT07PXIB9101209PR), Ministerio de Ciencia e Innovación (SAF 07-64873), Asociación Española contra el Cáncer, Fundación Científica and Junta de Barcelona, Fundación de Investigación Médica Mutua Madrilena, Fundación Privada Olga Torres and Fundación Barrié de la Maza.

Las grandes hazañas nunca son obra de una única persona, y esta tesis, aunque no es una gran hazaña, sí que contiene el trabajo y esfuerzo de otros, y no yo sola, que hoy merecen reconocimiento.

A Clara, por su confianza en mí a lo largo de estos años, por todas esas tardes de discusión sobre *odds ratios*, réplicas y CNVs. Porque gran parte de esta tesis es también suya.

A Ángel, porque gracias a él estoy hoy escribiendo estas palabras; por ser siempre el contrapunto de mi escepticismo *asociativo*.

To JB, for his neverending wisdom and infinite patience. To Ian, Luis and everyone else at the WTCHG; working with you has been both an honour and a pleasure.

A mis niños del Nesquik, los que lo son, y los que lo han sido, porque todas y cada una de las personas que alguna vez han subido a tomar el café conmigo han tenido que sufrir mis idas y venidas con los estudios de asociación, porque muchas de las grandes revelaciones sobre la Ciencia han surgido en conjunto con vosotros, y por haber compartido tantas risas/discusiones/búsquedas en la wikipedia/verdades universales. Al resto de la gente de la Fundación, que me ha sufrido/ayudado/apoyado durante estos casi 5 años.

A mis padres y mis hermanos, que me han querido a pesar de que decidiera estudiar biología. Por ayudarme a cumplir este sueño, y por ser siempre mi apoyo incondicional.

A Juan, consolador en ratos de tristeza y réplicas negativas y cómplice de mi felicidad en asociaciones positivas; por sacar lo mejor de mí y hacerme mejor persona; por ser *my other half*.

A todos los pacientes e integrantes de EPICOLON, porque gracias a su generosidad desinteresada, hoy esta tesis es posible.

A pai, mamá, Fiz, y Ate;

ó meu Xanciño

Science is not belief, but the will to find out

(Anonymous)

INDEX

BACKGROUND

Colorectal cancer (CRC) is one of the most frequent forms of neoplasia, being the 3rd most common cancer in men (663,000 cases per year, 10% of the total), and the 2nd in women (571 000 cases, 9.4% of the total). The global prevalence rate for CRC is 11.5%, but the distribution of the CRC burden is quite heterogeneous, with more than 60% of the cases occurring in developed regions[1].

1. TUMORIGENESIS

CRC is a late-onset disease, particularly prevalent in men (1.4:1 sex ratio), with and a median age at diagnosis of 72 years. Almost 90% of these cases will occur at ages of 50 or over, and 70% of these will be over 65[2]. This means that the development of CRC is a slow multistep process. A genetic model for CRC development, involving both somatic mutations and epigenetic changes, was first postulated by Fearon and Vogelstein in 1990[3]. Although this *chromosomal instability* route has been observed to be the cause of up to 85% of CRC tumours[4], it is not the only physiological mechanism that can give rise to CRC. It was later revealed that a considerable proportion of CRC (around 15%) arises through defects in the DNA mismatch repair system, leading to *microsatellite instability*[5] (Figure 1).

2. RISK FACTORS, DISEASE AETIOLOGY AND CRC GENETICS

CRC is considered a complex disease. This means that it arises as a result of the interplay between many genetic variants and environmental factors[6]. Low-fiber diets, red meat consumption, obesity, alcohol intake or smoking habits have been related to the development of the disease[7].

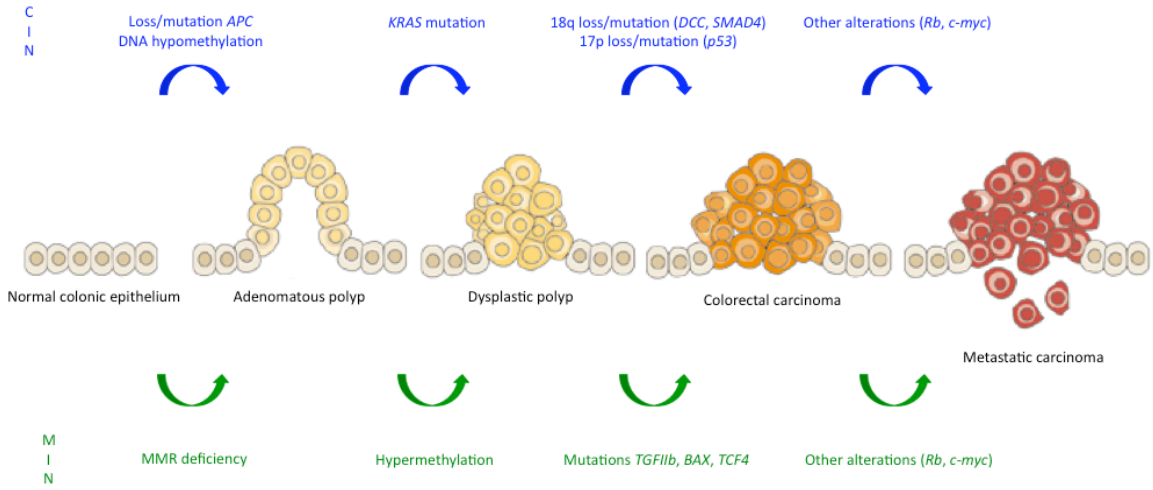


Figure 1. Genetic model for CRC tumorigenesis. CRC development may principally arise by two different paths: chromosomal instability (CIN-blue) or microsatellite instability (MIN-green). The CIN sequence usually starts with mutations in the tumour suppressor gene *APC* (which may be enhanced by defects on *MUTYH*[8]). This leads to genomic hypomethylation, an increased replication rate and a higher incidence of aneuploidies during cell division, causing *KRAS* mutations. This further enables adenoma growth and clonal expansion of the cells, with additional mutations in *DCC*, *SMAD4* and *p53* empowering the final transformation into a carcinoma[3]. Microsatellites are short DNA sequences highly prone to length variation due to their iterative nature. The accuracy in the replication of these DNA segments during cell division is ensured by the mismatch repair machinery (MMR), made up of the MLH1, MSH2, MSH6 and PMS2 proteins. Defects in these genes fuel replication errors in microsatellite locations throughout the genome, and many genes enriched for these sequences, such as *TFGIIb*, *BAX* or *TCF4*[9], are mutated. These ultimately enhance cell hyperproliferation and avoidance of apoptosis, subsequently initiating the CRC carcinogenetic sequence. Further alterations in other genes, such as *Rb* or *c-myc* have been proposed as the changes underlying the invasive potential of the tumour cells[10]. *Adapted from Knudson et al.*[11].

The three-way interaction: hereditary syndromes, familial and sporadic CRC

Genetic susceptibility is thought to explain a significant proportion of the incidence of complex diseases. This genetic portion is usually represented by a small Mendelian component determined by rare high-penetrance mutations, a middle-sized familial factor driven by the interaction of several common variants, each conferring a modest effect on disease risk, and a large sporadic fraction, mostly induced by environmental variables. The importance of these common low-penetrance variants was stated in the

Common Disease-Common Variant (CDCV) hypothesis, which postulated that the genetic component of common complex diseases was mostly due to variants of low/moderate effect that appeared at an elevated frequency in the population[12].

For CRC, twin studies have estimated that inherited predisposition might account for up to 35% of the cases[13]. Highly penetrant mutations have been described to underlie the hereditary CRC syndromes, namely Familial Adenomatous Polyposis (a result of mutations in the *APC* gene), *MUTYH*-associated polyposis (*MUTYH*), Lynch (*MLH1*, *MSH2*, *MSH6* and *PMS2*-MMR mutations) and Peutz-Jeghers syndromes (*LKB1/STK11)*, Juvenile Polyposis (*SMAD4* and *BMPR1A*)[14] and the Hereditary Mixed Polyposis syndrome (*BMPR1A*) [15]. The identification of the mutations leading to these syndromes has extensively relied on the use of linkage studies. With such penetrant effects, it is likely that the mutations causing these effects are very recent, and therefore, the chances of recombination in the corresponding haplotype will be very small, thus making it long. It is then appropriate to believe that markers flanking the mutations will co-segregate with the disease and so they could be used to identify these loci. Other potential loci connected to hereditary or familial phenotypes have been identified by linkage at 3q21-q24, 9q22.2-31.2[16], 7q31[17], 11q13.3, 14q24.2 and 22q[18], although the underlying genes responsible for the phenotype have not yet been found.

Association: from candidate genes to genome-wide association studies

Linkage analysis has achieved only limited success in the identification of CRC susceptibility loci. This makes sense, since highly penetrant mutations are only responsible for around 5% of the CRC cases[19]. For the familial and sporadic settings, which represent approximately 15% and 80% of the cases respectively, the genetic susceptibility is though to confer only low/moderate risk, and therefore linkage studies

show little power of detection. This is due to the fact that variants increasing CRC incidence in the expected range (1.5-2.0) will rarely cause multiple-case families and are therefore impossible to identify through linkage[20].

Association studies have been postulated as the most reasonable strategy in the identification of common modest-risk variants. The typical association designs are case-control studies, in which the frequency of the potential susceptibility variant is compared between a group of affected and healthy individuals. The use of this approach was greatly encouraged by the completion of the Human Genome Project (HGP http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml), and the discovery of a high genomic variation in the form of single nucleotide polymorphisms (SNPs). SNPs are the most abundant genetic markers in the genome (over 12 million), constituting a major source of inter-individual genetic and phenotypic variation. The construction of large SNP maps and databases after the completion of the HGP boosted the use of association studies in the discovery of new susceptibility variants for CRC and other diseases[21].

Initially, the strategy used to scan for new susceptibility variants was principally a *direct candidate-gene* approach, which relied on the evaluation of specific potentially relevant SNPs (mainly non-synonymous and regulatory changes) within selected genes that were thought to be important in the development of the disease. The main advantage this strategy confers is the easy biological interpretation of the associations, since changes in these genes can be easily linked to the neoplasic process. Soon whole genetic pathways, genes located in previously determined regions of linkage were extensively evaluated. For CRC, several approaches screening particular genes such as *APC*[22], carcinogenesis-related pathways, like *Wnt* and DNA-repair ones[23] or the mouse *PTPRJ* candidate [24] were thoroughly examined.

Nonetheless, no common variants contributing to colorectal cancer risk could be identified and consistently replicated, and the analysis through candidate-gene strategies proved largely insufficient to characterise the whole of the genetic variation underlying most diseases, with CRC being no exception. This was probably due to a number of reasons: firstly, candidate-gene studies were restricted to potentially relevant loci based on *a priori* estimates of the biological mechanisms. Secondly, most of these early designs were clearly underpowered to detect loci with the expected risk effect under the CDCV hypothesis. Moreover, liberal thresholds were used to call positive association findings and this usually resulted in the lack of replication for most of these hits[25].

Luckily, the completion of the HGP supplied with yet another important tool: the possibility to study the fine resolution of linkage disequilibrium (LD) in the genome. LD had been already described in the 60s as the non-random association of alleles at multiple loci in a proportion greater than the expected by Mendelian law[26]. The discovery of LD implicated that long DNA segments were transmitted together for generations, generating haplotype blocks in which the genotypes of the different SNPs are correlated with one another. The Haplotype Map (or HapMap) Project has been the key to the evaluation of these LD patterns in the genomic level[21]. The knowledge in the distribution of these blocks allows for the screening of the genetic variation in extensive regions by genotyping a relatively reduced set of informative markers, or *tagSNPs*. The implementation of this *indirect* approach (*versus* the *direct* approach usually undertaken by candidate-gene studies, in which the genotyped variant is thought to be the functional cause of the disease susceptibility), coupled to the progression of high-throughput technologies and a reduction in genotyping costs, enabled association studies to be performed on a genome-wide basis, giving rise to the so-called *Genome-Wide Association Studies* (GWAS).

GWAS represent a comprehensive yet unbiased option for association strategies, since they make no prior assumption on the location or functionality of the variants determining CRC risk. They represent an extremely good means of identifying common SNPs with modest effects on genotype. Since their implementation in late 2007, they have resulted in the discovery of several susceptibility loci in a variety of complex diseases and quantitative traits[27].

For CRC, GWAS have successfully identified 16 SNPs in 14 risk loci: rs6983267 at 8q24[28-31], rs4939827 at 18q21.1[32], rs4779584 at 15q13.3[33], rs3802842 at 11q23.1[31], rs16892766 at 8q23.3, rs10795668 at 10p14[34], rs4444235 at 14q22.2, rs9929218 at 16q22.1, rs10411210 at 19q13, rs961253 at 20p12.3[35], rs6691170 and rs6687758 at 1q41, rs10936599 at 3q26.2, rs11169552 and rs7136702 at 12q13.13 and rs4925386 at 20q13.33[36]. As expected by the complex disease model, these identified associations have all modest effects on disease risk, with odds ratios typically below 1.5.

Even when GWAS have been quite successful for the discovery phase, we must bare in mind that the identified variants are most of the times not the functional ones. This means that the definition of the molecular mechanisms through which they influence disease risk and/or phenotype is yet to be assured for most of the cases. It is however quite outstanding, that some of these SNPs nearby genes seem to belong to the Transforming Growth Factor-beta (TGF-ß) signalling pathway, which has been extensively related to colorectal carcinogenesis[37]. On the other hand, these loci altogether explain only around 7% of the excess genetic susceptibility to CRC[36]. This highlights the need for further collaborative efforts involving larger sample sets and the combination of GWAS data that will hopefully lead to the identification of new variants that can explain the remaining proportion of expected genetic susceptibility. A summary of the genetic susceptibility loci already identified is shown on Figure 2.

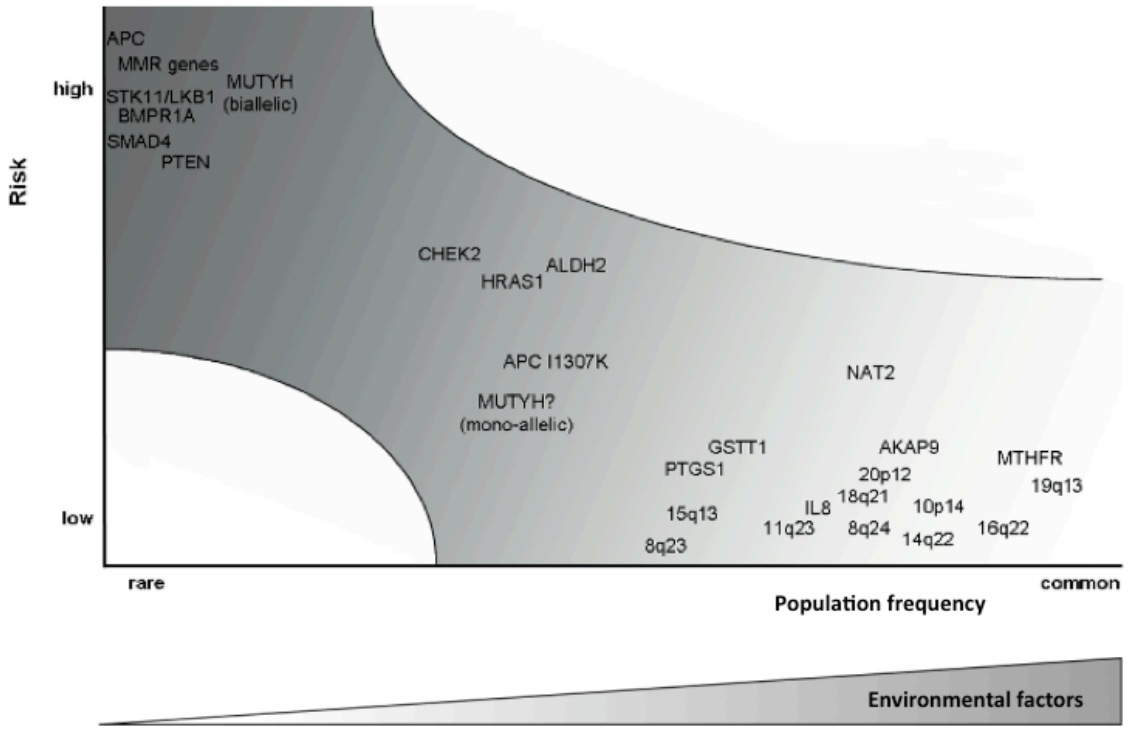


Figure 2. The genetic risk factors in CRC. Overview of the loci related to CRC susceptibility and development. Frequent variants conferring high risk have not been found, since they are likely to affect the individual´s viability, whereas rare variants with low frequencies are difficult to detect. High-penetrance rare mutations are typical of the hereditary syndromes, whereas more frequent moderately penetrant variants, discovered by association studies, are though to play an important role in the development of the disease. *Adapted from A. Middeldorp; Genetics and tumor genomics in familial colorectal cancer 2010.*

Copy-number variation

The spectrum of human genetic variation ranges from single base pair changes to large chromosomal rearrangements. For much of the last decade, aims at explaining the genetic susceptibility underlying common diseases were focused on the inspection of SNPs. However, several independent reports confirmed in the last few years that there is indeed another form of variation that may involve an equally great proportion of the human genome: submicroscopic structural variation, also known as copy-number variants (CNVs)[38,39].

CNVs are events that range from 1kb to the microscopic detectable limit (~3Mb), and include duplications, deletions and inversions of DNA segments. Approximately 80% of these CNVs segregate at allele frequencies of 5% or greater in the population, making them a common source of genetic variation. Although the incidence of CNVs is considerably smaller than that of SNPs, they can affect large stretches of DNA, adding up to an estimated 12% of the genome[40]. Several CNV maps have been performed so far to establish the distribution of these events in the genome[41-43]. Most of the information gathered so far on CNVs is available at the Database of Genomic Variants (http://projects.tcag.ca/variation/) and the UCSC browser (http://genome.ucsc.edu/cgi-bin/hgGateway).

It has been observed that the incidence of CNVs appears to be greater in regions with segmental duplications (SD)[41]. SDs are segments of DNA over 1kb in size that have 90% sequence homology with other locations elsewhere in the genome. The presence of these duplications could lead to non-allelic homologous recombination (NAHR) events between copies and the eventual creation of CNVs. NAHR was the first mechanism proposed for CNV creation, but processes such as double-strand break repair, B-DNA structures or transposition of mobile elements have also been seen to contribute to CNV formation[44]. On the contrary to what might have been expected, CNVs do not seem to be biased against genic regions, and around 40% of CNV events have been seen to overlap both RefSeq and OMIM genes[45]. Amongst these, there is a noticeable enrichment for genes with roles in adaptability and fitness, particularly olfactory receptors and genes implicated in xenobiotic metabolism. This could implicate structural variation in the dynamics of response to external stimuli and thus gene and organismal adaptation and evolution[46].

It has also been postulated that CNVs may play an important role in the genetics of complex diseases[42]. Although most of these variants are presumed to be benign, they can have subtle influences on phenotypes and disease development, as has been described for HIV-1 infection[47] and glomerulonephritis[48]. It is thus of great importance to test the possibility that CNVs may confer susceptibility to other diseases, such as CRC. It has been suggested that most common CNVs (around 77%) are in linkage disequilibrium with SNPs[45,49,50]. Hence, the implications of most of these in disease aetiology could be effectively evaluated indirectly through SNP analysis. However, there are two main reasons to justify the use of CNV-specific genotyping procedures. Firstly, CNVs appear to be present at low-coverage regions in both the reference human genome sequence and HapMap, probably because of the problematics in the sequencing of these regions. This implies that the coverage of the available SNP genotyping arrays is actually biased against CNV locations. Secondly, the presence of CNV events may generate inconsistencies that would make them unlikely to be accurately assessed in SNP-based studies. For instance, it has been described that deletions are prone to creating Mendelian errors, whereas duplications tend to leave high residual missing genotype rates and fail Hardy-Weinberg equilibrium (HWE)[43]. Fortunately, the new generations of arrays have accounted for these problems, and there are now a variety of CNV (Agilent Technologies) and dual SNP-CNV chips (Affymetrix 6.0, Illumina 1M) that specifically target the CNV component independently. A variety of algorithms have been proposed to infer copy-number status from the probe intensities obtained in these arrays (QuantiSNP[51], Birdsuite[52], PennCNV[53]), in order to enable later association measurements.

The need to explore the genome for other sources of variability that could explain the remaining of the genetic component of common disorders has been increasingly urgent

in the past years. The first GWAS studies seemed to point that CNV contribution to common disease had been overestimated, for most of the genetic risk conferred by these variants had already been accounted for in SNP studies[45,50]. Nevertheless, there have also been some encouraging findings, particularly in the field of neuropsychiatric disorders that have shown that CNVs can still explain part of the remaining unexplained genetic variation in common diseases[54].

3. CRC MANAGEMENT AND TREATMENT

The degree of CRC development and thus the therapeutic strategy to follow is determined by tumour staging. Historically, Duke´s staging system has been used for CRC diagnosis, although in the last few years there has been a trend towards an integration with the Tumour-Node-Metastasis (TNM) classification for solid tumours[55]. Although surgery is the most common choice of action for all-stage CRC patients, the use of chemotherapy is commonly indicated as well.

Additionally, there is significant divergence in treatment strategies depending on the location of the tumour. This is principally due to the anatomical differences between colon and rectum, and the consequent implications these have over surgical resection of the tumour and its aftermath. It is for this reason that colon and rectal cancer strategies will be discussed separately.

Colon cancer

As has been stated in the former paragraph, surgery is the cornerstone treatment for colon cancer, particularly when it is localised (stages I, II and III). Moreover, the use of adjuvant (postoperative) treatments is also common for node-positive (stage III) disease

patients, for whom recurrence is a serious problem[56]. Adjuvancy has also been tested for stage II patients in clinical trials, although its use remains controversial, for the shown improvements in overall survival rates have been outweighed by the treatment toxicities and comorbidities[57]. The usual drug choices for these stages are typically 5-fluorouracil (5-FU) and leucovorin (LV) in monotherapy, or in combination with oxaliplatin (FOLFOX). The mechanisms of action of these agents will be explained in the coming sections (See *Pharmacogenetics: Five-fluorouracil* and *Oxaliplatin*). On the other hand, chemotherapy is also extensively used as a palliative measure for metastatic colon cancer (stage IV). Treatment at this stage includes administration of 5-FU/LV in combination with oxaliplatin (FOLFOX) or irinotecan (FOLFIRI)[58] (see *Pharmacogenetics* section). The use of monoclonal antibodies targeting the vascular endothelial growth factor (VEGF) and the epidermal growth factor receptor (EGFR) is also common at this stage. VEGF inhibitors, such as bevacizumab, are used to prevent the formation of new blood vessels, which would enable tumour growth and spreading, whereas two humanised monoclonal antibodies are used for the inhibition of EGFR-mediated signalling: cetuximab and panitumumab[59].

Rectal cancer

The management of rectal cancer varies somewhat from that of the colon. Radical surgery is also the main treatment of choice, but is limited by the frequent impossibility of achieving wide margins due to the presence of the bony pelvis. Therefore, the risk of local recurrence is much higher than in colon cancers[60]. Neoadjuvant (preoperative) 5-FU/LV chemotherapy along with radiotherapy (*chemoradiation*), followed by adjuvant chemotherapy with 5-FU/LV or FOLFOX is considered as the standard protocol for

rectal cancer stages II and III[58]. For rectal stage IV cancer patients, same protocol as for colon stage IV applies (FOLFOX and FOLFIRI).

4. PHARMACOGENETICS

It has been long known that there is a high inter-individual diversity in the outcome of drug administration. These differences may be due to multiple factors, such as health state of the patient, sex, age or co-administration with other drugs. However, the observation during the 1940s that some of these unusual drug responses presented familial clustering resulted in the realisation that at least part of this variation may also be due to genetic factors[61]. This led to the birth of *pharmacogenetics*, as the scientific field that aims to understand the genetic basis of this observed variability in therapeutic outcome in order to individualise treatments for improved response and reduced toxicity. Ever since this discovery, many studies have aimed at unravelling the genetic contributions to drug therapy outcome.

Why these variations? Pharmacokinetics and pharmacodynamics

The genetics underlying drug efficacy and toxicity may implicate variants in many genes. Then it is feasible to assume that many of the genes that encode for proteins involved in drug availability or enabling of drug function may influence therapeutic results.

Pharmacokinetics has been used in relation to the differences in the delivery of a drug or metabolite to its target molecules. These discrepancies may arise in processes such as drug absorption, distribution, metabolism or elimination[62]. Several variants have already

been described that contribute to variability in pathways of drug disposition. The human cytochrome P450 (*CYP*) gene family members may be a good example of this[63].

Likewise, the term pharmacodynamics describes the relationship between drug concentration and its effect. The appearance of pharmacodynamic variation may arise directly from variability within the drug target or from the genetic variation in other molecules with which the drug or the target interact[62].

The candidate-gene approach in pharmacogenetics

Early pharmacogenetic studies greatly focused on variants with Mendelian effects on response. By these means, a handful of mutations have been discovered that account for a large proportion of the population effects[64]. Nevertheless, given the normal distributions observed in the phenotypes of drug administration, both for efficacy of response and the development of toxicities, it is quite rare that single polymorphisms will explain the whole variety of outcomes observed. There is also the fact that most of these high-risk genetic variants are uncommon, and therefore may only explain a small proportion of the population variances. Thus, it is widely believed that most drug effects may be, same as for complex diseases, polygenic in nature[65].

As in disease association studies, the classical approach to identifying new susceptibility loci has been the analysis of single-gene variants. In this sense, pharmacokinetic-related genes, such as drug metabolisers and drug transporters, and pharmacodynamic effectors, such as drug targets, have been screened for evidences of association[66,67]. Extensive studies have also been performed to screen the genetic variation in entire biological and pharmacological pathways, such as the DNA-repair machinery genes (BER, NER) [68] or the genes involved in folate metabolism[69].

Pharmacogenetics in cancer

The matching of patients to the protocol most likely to be effective and less harmful is of essential importance in cancer chemotherapy, since many anticancer drugs have narrow therapeutic indexes and the threshold delimiting the therapeutic range and the toxic response is diffuse. It is for this reason that there have been extensive studies on the pharmacogenetics of anticancer drugs. A brief description of each of the most common CRC chemotherapy drugs and the associations found will be thus presented.

Five-fluorouracil (5-FU)

Five-fluorouracil was introduced about 50 years ago as the first systemic chemotherapy agent for CRC treatment[70]. Since then, it has formed the backbone of treatment for both adjuvant and advanced settings in CRC chemotherapy, particularly in combination with its cofactor leucovorin (LV). Capecitabine is an oral pro-drug of 5-FU. This molecule is a frequent alternative to 5-FU administration, since it effectively increases the concentration of 5-FU in the neoplasic cells[71]. The mechanism of action of these fluoropyrimidine compounds relies on the inhibition of thymidylate synthase (TYMS), the rate-limiting enzyme in pyrimidine nucleotide synthesis[72] (Figure 3).

The administration of 5-FU-LV/capecitabine has however some potentially serious side effects. Gastrointestinal and haematological adverse drug reactions (ADRs), and hand-foot syndrome are frequently observed[73]. It has been estimated that 26-65% of these variations in susceptibility to 5-FU-induced toxicity are due to genetic components. Several variants have been described as associated with this inherited predisposition, and all of them have been identified within important genes related to 5-FU metabolism (Table 1).

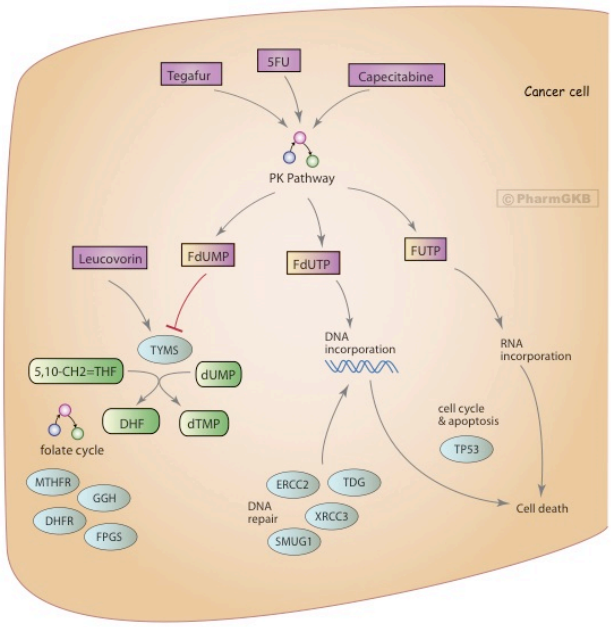


Figure 3. Fluoropyrimidine mechanism of action. Fluoropyrimidines (5-FU and capecitabine) are primarily converted to fluorodeoxyuridine diphosphate (FdUDP). FdUDP can either be phosphorylated or dephosphorylated to generate the active metabolites fluorodeoxyuridine triphosphate (FdUTP) and fluorodeoxyuridine monophosphate (FdUMP), respectively. FdUMP inhibits TYMS, whereas FdUTP is directly incorporated into growing DNA chains, resulting in impaired strand elongation and an increase in DNA fragmentation. The collateral formation of fluorouridine triphosphate (FUTP) also contributes to the toxic effects of 5-FU by disruption of mRNA function[93]. The detoxification of the fluoropyrimidines occurs almost exclusively by the action of the dihydropyrimidine dehydrogenase (DPYD) enzyme in the liver[94]. *Adapted from Klein et al. 2001*[95] *(The Pharmacogenomics Journal); copyright of PharmGKB.*

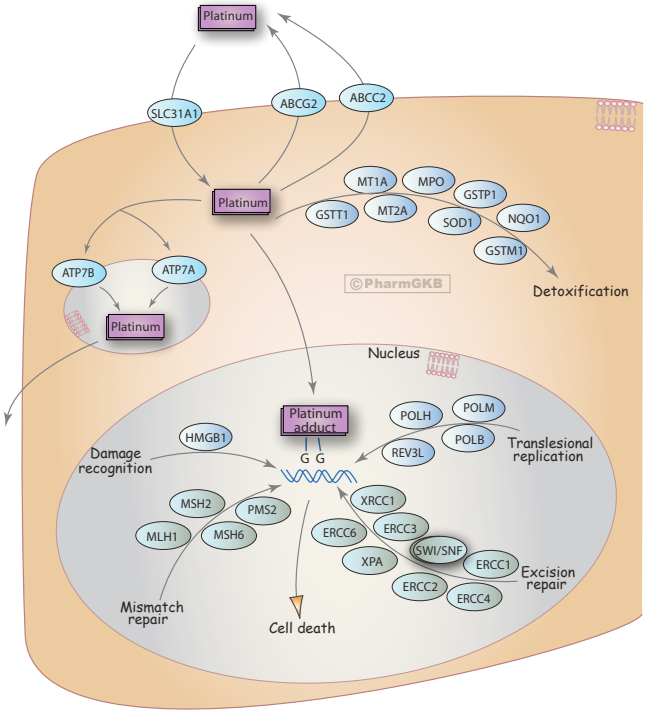


Figure 4: Mechanism of action of platinum anticancer drugs. Platinum compounds are able to destroy cancerous cells by interfering with DNA, via inter- and intra-strand crosslinks, and DNA-protein binding, thereby preventing cell division and growth. The formation of these DNA adducts may be overcome by several cellular mechanisms: the cell´s system repair genes (both the mismatch and the excision repair), the translesional replication machinery (POL and REV3L proteins) or the damage recognition system (HMGB1). Detoxification of platinum may also occur through glutathione-S-transferases (GSTs); members of this gene family have been implicated in both detoxification events and carcinogenesis[98]. *Adapted from Marsh et al. 2009 Pharmacogenetics Genomics; copyright of PharmGKB.*

Oxaliplatin

Oxaliplatin is a third-generation platinum compound that induces the formation of DNA adducts, inducing an apoptotic response in the tumour cell[78] (Figure 4). It has modest activity against colorectal cancer by itself, but is shown to produce clinical benefits in response rates and overall survival in addition to the classic 5-FU-LV (a combination typically known as FOLFOX)[79]. Because of this, the use of FOLFOX has been widely implemented for adjuvant and palliative cancer therapies.

As happens with other chemotherapeutical agents, administration of oxaliplatin/FOLFOX has considerable side effects. A severe cumulative sensory neuropathy that may endure even after treatment ending has been the most substantial of these[80]. Diarrhoea, neutropenia and nausea/vomiting may also appear with high prevalences[81]. This heterogeneity in treatment outcome has also been targeted by pharmacogenetic studies. The most remarkable associations are shown on Table 1.

Table 1: Description of the variants associated with drug outcome in CRC treatments.

TREATMENT	GENE	FUNCTION	CHANGE*	MECHANISM OF ACTION	RELATED PHENOTYPE	REF
5-FU	*DPYD*	5-FU detoxification	IVS14+1 (rs3918290)	Exon 14 skipping; protein truncation	Grade IV neutropenia	82
	TYMS	Pyrimidine synthesis	5´UTR VNTR (2R or 3R) (rs34743033) 3´UTR 6bp indel (rs34489327) G>C transversion at the 2nd repeat of 3R (no rs found)	3RG+6bp insertion haplotype increases TYMS activity	Lower toxicity	83
	MTHFR	Folate metabolism	c.C677T (rs1801133)	T allele	Better response; higher toxicities	84
Oxaliplatin	*GSTP1*	Oxaliplatin detoxification	p.I105V (rs1695)	Unknown	Reduced enzymatic activity, higher toxicity	85
Irinotecan	*UGT1A1*	Irinotecan detoxification	UGT1A1*28; Dinucleotidic expansion (6-7 repeats) (rs8175347)	Impairs union of TFIID with gene promoter	Neutropenia and diarrhoea	86
			UGT1A1*6; G71R (rs4148323)	Slower degradation rate	Neutropenia in Asian populations	87
Cetuximab	*KRAS*	Proto-oncogene	p.G12D (no rs found)	KRAS activation enables cell proliferation without EGFR induction	Poor response to cetuximab treatment	88

**Denotes both dbSNP and common literature codings for this variant.*

Other treatments: irinotecan and targeted therapies

Irinotecan is a topoisomerase 1 inhibitor used in solid tumour chemotherapy. This impairs DNA replication, resulting in cell death. As with oxaliplatin, this drug is typically administered in combination with 5-FU-LV (commonly referred to as FOLFIRI)[89]. Around 30% of patients administered with irinotecan develop severe neutropenia or diarrhoea during the treatment. A polymorphism in the *UGT1A1* gene, responsible for irinotecan detoxification through glucuronidation has been linked to these adverse effects[86] (Table 1).

Treatment with cetuximab, a monoclonal antibody directed against EGFR, improves overall and progression-free survival in patients with colorectal cancer that have not responded to chemotherapy. However, the effectiveness of cetuximab administration depends heavily on the mutational status of the *KRAS* gene. Is has been observed that *KRAS* mutated tumours do not benefit from cetuximab administrations[88].

Pharmacogenomics

Although candidate-gene approaches have had reasonable success in identifying genetic variants that are important in specific phenotypes, the evaluation of a gene in isolation will most likely never provide the sensitivity and specificity that is needed for tailored treatment decisions. This evidences an urgent need to detect some new polymorphisms that are able to predict a bigger portion of the expected heritability.

Along these lines, pharmacogenetics may as well follow the path of disease genetics, and take advantage of the knowledge acquired on the genome over the last two decades and the developments on genotyping technologies. The term pharmacogenomics has been described as the wide-range genomic application of pharmacogenetics, although several definitions with different connotations are available[90]. Pharmacogenomics could

mean a new chapter for variant discovery, and an opportunity to find new understanding on the molecular basis of drug disposition and drug action without the constraints of predetermined candidate genes.

Since 2007, several GWAS have appeared on the pharmacogenetic field, analysing drug outcomes in different diseases, such as asthma, psychiatric disorders or cardiovascular disease[91]. Less than half of these GWAS published on response have however yielded significant results. Positive findings in this field include the identification of the genes responsible to variation in response to coumarin anticoagulants. This effect has been mainly linked to variation in only three principal genes: *CYP2C9*, *VKORC1* and *CYP4F2* explaining almost 64% of the total attributable genetic variation[92]. Fewer GWAS have been published on ADRs (around 30% of the total GWAS), and only two of them have successfully reported significant findings: fluoxacin-induced liver injury linked to *HLA-B*5701*[93] and simvastatin-induced myopathy, determined by variants in the *SLCO1B1* gene[94].

Nevertheless, the number of GWAS in pharmacogenetics is increasing rapidly. Although most of these studies have failed in identifying any new variants, they have pointed out at some new loci that could be very interesting for follow-up. Further GWAS studies in larger cohorts could verify the importance of these associations and their potential application to tailored drug therapies.

5. REFERENCES

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010;127:2893-917.
2. Parker SL, Tong T, Bolden S, Wingo PA. Cancer statistics, 1996. CA Cancer J Clin 1996;46:5-27.
3. Fearon ER, Vogelstein B. A genetic model for colorectal tumorigenesis. Cell 1990;61:759-67.
4. Powell SM, Zilz N, Beazer-Barclay Y, et al. APC mutations occur early during colorectal tumorigenesis. Nature 1992;359:235-7.
5. Henry T. Lynch, M.D., and Albert de la Chapelle. Hereditary colorectal cancer. 2003;.
6. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 2002;31:33-6.
7. Wei EK, Wolin KY, Colditz GA. Time course of risk factors in cancer etiology and progression. J Clin Oncol 2010;28:4052-7.
8. Al-Tassan N, Chmiel NH, Maynard J, et al. Inherited variants of *MYH* associated with somatic G:C-T:A mutations in colorectal tumors. Nat Genet 2002;30:227-32.
9. Peltomaki P. Deficient DNA mismatch repair: a common etiologic factor for colon cancer. Hum Mol Genet 2001;10:735-40.
10. Karakosta A, Golias C, Charalabopoulos A, Peschos D, Batistatou A, Charalabopoulos K. Genetic models of human cancer as a multistep process. Paradigm models of colorectal cancer, breast cancer, and chronic myelogenous and acute lymphoblastic leukaemia. J Exp Clin Cancer Res 2005;24:505-14.
11. Knudson AG. Two genetic hits (more or less) to cancer. Nat Rev Cancer 2001;1:157-62.
12. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001;17:502-10.
13. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer-analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78-85.
14. Lynch HT, de la Chapelle A. Hereditary colorectal cancer. N Engl J Med 2003;348:919-32.
15. Cao X, Eu K, Kumarasinghe M, *BMPR1A* Li H, Loi C, Cheah P. Mapping of hereditary mixed polyposis syndrome (HMPS) to chromosome 10q23 by genomewide high-density single nucleotide polymorphism (SNP) scan and identification of loss of function. J Med Genet 2006;43:e13.
16. Kemp Z, Carvajal-Carmona L, Spain S, et al. Evidence for a colorectal cancer susceptibility locus on chromosome 3q21-q24 from a high-density SNP genome-wide linkage scan. Hum Mol Genet 2006;15:2903-10.
17. Neklason DW, Kerber RA, Nilson DB, et al. Common familial colorectal cancer linked to chromosome 7q31: a genome-wide analysis. Cancer Res 2008;68:8993-7.
18. Djureinovic T, Skoglund J, Vandrovcova J, et al. A genome wide linkage analysis in Swedish families with hereditary non-familial adenomatous polyposis/non-hereditary non-polyposis colorectal cancer. Gut 2006;55:362-6.
19. Houlston RS, Peto J. The search for low-penetrance cancer susceptibility alleles. Oncogene 2004;23:6471-6.

20. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science 1996;273:1516-7.
21. International HapMap Consortium, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851-61.
22. Picelli S, Zajac P, Zhou XL, et al. Common variants in human CRC genes as low-risk alleles. Eur J Cancer 2010.
23. Webb EL, Rudd MF, Sellick GS, et al. Search for low penetrance alleles for colorectal cancer through a scan of 1467 non-synonymous SNPs in 2575 cases and 2707 controls with validation by kin-cohort analysis of 14 704 first-degree relatives. Hum Mol Genet 2006;15:3263-71.
24. Toland AE, Rozek LS, Presswala S, Rennert G, Gruber SB. PTPRJ Haplotypes and Colorectal Cancer Risk. Cancer Epidemiol Biomarkers Prev 2008;17:2782-5.
25. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med 2002;4:45-61.
26. Lewontin R, Kojima K. The evolutionary dynamics of complex polymorphisms. Evolution 1960;14:458-72.
27. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.
28. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 2007;39:984-8.
29. Zanke BW, Greenwood CM, Rangrej J, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 2007;39:989-94.
30. Haiman CA, Patterson N, Freedman ML, et al. Multiple regions within 8q24 independently affect risk for prostate cancer. Nat Genet 2007;39:638-44.
31. Tenesa A, Farrington SM, Prendergast JG, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 2008.
32. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 2007;39:1315-7.
33. Jaeger E, Webb E, Howarth K, et al. Common genetic variants at the *CRAC1* (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet 2008;40:26-8.
34. Tomlinson IP, Webb E, Carvajal-Carmona L, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 2008.
35. Houlston RS, Webb E, Broderick P, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40:1426-35.
36. Houlston RS, Cheadle J, Dobbins SE, 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;42:973-7.
37. Hemminki K, Lorenzo Bermejo J, Forsti A. The balance between heritable and environmental aetiology of human disease. Nat Rev Genet 2006;7:958-65.

38. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949-51.
39. Sebat J, Lakshmi B, Troge J, et al. Large-scale copy number polymorphism in the human genome. Science 2004;305:525-8.
40. Eichler EE, Nickerson DA, Altshuler D, et al. Completing the map of human genetic variation: A plan to identify and integrate normal structural variation into the human genome sequence. Nature 2007;447:161.
41. Tuzun E, Sharp AJ, Bailey JA, et al. Fine-scale structural variation of the human genome. Nat Genet 2005;37:727-32.
42. Conrad DF, Andrews TD, Carter NP, Hurles ME, Pritchard JK. A high-resolution survey of deletion polymorphism in the human genome. Nat Genet 2006;38:75-81.
43. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. Nature 2006;444:444-54.
44. Hastings P, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nature Reviews Genetics 2009;10:551-64.
45. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. Nature 2009;464(7289):704-12.
46. Feuk L, Carson AR, Scherer SW. Structural variation in the human genome. Nature Reviews Genetics 2006;7:85-97.
47. Gonzalez E, Kulkarni H, Bolivar H, et al. The influence of *CCL3L1* gene-containing segmental duplications on HIV-1/AIDS susceptibility. Science 2005;307:1434.
48. Aitman TJ, Dong R, Vyse TJ, et al. Copy number polymorphism in *Fcgr3* predisposes to glomerulonephritis in rats and humans. Nature 2006;439:851-5.
49. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 2006;38:82-5.
50. Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010;464:713-20.
51. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35:2013-25.
52. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008;40:1253-60.
53. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007;17:1665-74.
54. Need AC, Ge D, Weale ME, et al. A genome-wide investigation of SNPs and CNVs in schizophrenia. PLoS Genet 2009;5:e1000373.
55. Hutter RVP. At Last-Worldwide Agreement on the Staging of Cancer: Presidential Address. Archives of Surgery 1987;122:1235.
56. Cunningham D, Atkin W, Lenz HJ, et al. Colorectal cancer. Lancet 2010;375:1030-47.

57. Quasar Collaborative G, Gray R, Barnwell J, et al. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. Lancet 2007;370:2020-9.
58. Meyerhardt JA, Mayer RJ. Systemic therapy for colorectal cancer. N Engl J Med 2005;352:476-87.
59. Sawyers C. Targeted cancer therapy. Nature 2004;432:294-7.
60. Phillips R, Hittinger R, Blesovsky L, Fry J, Fielding L. Local recurrence following 'curative'surgery for large bowel cancer: I. The overall picture. Br J Surg 1984;71:12-6.
61. Beutler E, Dern RJ, Flanagan CL, Alving AS. The hemolytic effect of primaquine. VII. Biochemical studies of drug-sensitive erythrocytes. J Lab Clin Med 1955;45:286-95.
62. Roden DM, George Jr AL. The genetic basis of variability in drug responses. Nature Reviews Drug Discovery 2002;1:37-44.
63. Kim RB. Pharmacogenetics of CYP enzymes and drug transporters: remarkable recent advances. Adv Drug Deliv Rev 2002;54:1241-2.
64. Ingelman-Sundberg M. Genetic polymorphisms of cytochrome P450 2D6 (*CYP2D6*): clinical consequences, evolutionary aspects and functional diversity. The pharmacogenomics journal 2004;5:6-13.
65. Crowley JJ, Sullivan PF, McLeod HL. Pharmacogenomic genome-wide association studies: lessons learned thus far. Pharmacogenomics 2009;10:161-3.
66. Woosley RL, Chen Y, Freiman JP, Gillis RA. Mechanism of the cardiotoxic actions of terfenadine. JAMA: The Journal of the American Medical Association 1993;269:1532.
67. Tsunoda A, Nakao K, Watanabe M, Matsui N, Ooyama A, Kusano M. Associations of various gene polymorphisms with toxicity in colorectal cancer patients receiving oral uracil and tegafur plus leucovorin: a prospective study. Annals of Oncology 2011;22:355.
68. Olaussen KA, Dunant A, Fouret P, et al. DNA repair by ERCC1 in non–small-cell lung cancer and cisplatin-based adjuvant chemotherapy. N Engl J Med 2006;355:983-91.
69. Robien K, Boynton A, Ulrich CM. Pharmacogenetics of folate-related drug targets in cancer treatment. Pharmacogenomics 2005;6:673-89.
70. Heidelberger C, Ansfield FJ. Experimental and clinical use of fluorinated pyrimidines in cancer chemotherapy. Cancer Res 1963;23:1226.
71. Saif MW, Katirtzoglou NA, Syrigos KN. Capecitabine: an overview of the side effects and their management. Anticancer Drugs 2008;19:447-64.
72. Sobrero A, Guglielmi A, Grossi F, Puglisi F, Aschele C. Mechanism of action of fluoropyrimidines: relevance to the new developments in colorectal cancer chemotherapy. 2000;27:72.
73. Eng C. Toxic effects and their management: daily clinical challenges in the treatment of colorectal cancer. Nat Rev Clin Oncol 2009;6:207-18.
74. Parker WB, Cheng YC. Metabolism and mechanism of action of 5-fluorouracil. Pharmacol Ther 1990;48:381-95.
75. Heggie GD, Sommadossi JP, Cross DS, Huster WJ, Diasio RB. Clinical pharmacokinetics of 5-fluorouracil and its metabolites in plasma, urine, and bile. Cancer Res 1987;47:2203.
76. Klein T, Chang J, Cho M, et al. Integrating genotype and phenotype information: an overview of the PharmGKB project. The Pharmacogenomics Journal 2001;1:167-70.

77. Marsh S, McLeod H, Dolan E, et al. Platinum pathway. Pharmacogenetics and Genomics 2009;19:563.
78. Raymond E, Faivre S, Woynarowski JM, Chaney SG. Oxaliplatin: mechanism of action and antineoplastic activity. 1998;25:4.
79. de Gramont A, Figer A, Seymour M, et al. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. Journal of Clinical Oncology 2000;18:2938.
80. Cersosimo RJ. Oxaliplatin-associated neuropathy: a review. Ann Pharmacother 2005;39:128.
81. Haller D. Safety of oxaliplatin in the treatment of colorectal cancer. Oncology-Huntington 2000;14:15-20.
82. Van Kuilenburg ABP, Meinsma R, Zoetekouw L, Van Gennip AH. High prevalence of the IVS14 1G> A mutation in the dihydropyrimidine dehydrogenase gene of patients with severe 5-fluorouracil-associated toxicity. Pharmacogenetics and Genomics 2002;12:555.
83. Ichikawa W, Takahashi T, Suto K, Sasaki Y, Hirayama R. Orotate phosphoribosyltransferase gene polymorphism predicts toxicity in patients treated with bolus 5-fluorouracil regimen. Clinical cancer research 2006;12:3928.
84. Sohn KJ, Croxford R, Yates Z, Lucock M, Kim YI. Effect of the methylenetetrahydrofolate reductase C677T polymorphism on chemosensitivity of colon and breast cancer cells to 5-fluorouracil and methotrexate. J Nat Cancer Inst 2004;96:2:134.
85. Lecomte T, Landi B, Beaune P, Laurent-Puig P, Loriot MA. Glutathione S-transferase P1 polymorphism (Ile105Val) predicts cumulative neuropathy in patients receiving oxaliplatin-based chemotherapy. Clinical cancer research 2006;12:3050.
86. Iyer L, Das S, Janisch L, et al. *UGT1A1**28 polymorphism as a determinant of irinotecan disposition and toxicity. The Pharmacogenomics Journal 2002;2:43-7.
87. Jinno H, Tanaka-Kagawa T, Hanioka N, et al. Glucuronidation of 7-ethyl-10-hydroxycamptothecin (SN-38), an active metabolite of irinotecan (CPT-11), by human *UGT1A1* variants, G71R, P229Q, and Y486D. Drug Metab Disposition 2003;31:108.
88. Lievre A, Bachet JB, Le Corre D, et al. *KRAS* mutation status is predictive of response to cetuximab therapy in colorectal cancer. Cancer Res 2006;66:3992.
89. Douillard JY, Cunningham D, Roth A, et al. Irinotecan combined with fluorouracil compared with fluorouracil alone as first-line treatment for metastatic colorectal cancer: a multicentre randomised trial. The Lancet 2000;355:1041-7.
90. Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. Nature Reviews Genetics 2003;4:937-47.
91. Daly AK. Genome-wide association studies in pharmacogenomics. Nat Rev Genet 2010;11:241-6.
92. Cooper GM, Johnson JA, Langaee TY, et al. A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood 2008;112:1022.
93. Daly AK, Donaldson PT, Bhatnagar P, et al. *HLA-B*5701* genotype is a major determinant of drug-induced liver injury due to flucloxacillin**.** Nat Genet 2009;41:816-9.
94. Wilhelmsen L, Link E, Parish S, et al. *SLCO1B1* variants and statin-induced myopathy-a genomewide study. N Engl J Med 2008;359:789-99

AIMS AND OUTLINE

Heritability in colorectal cancer (CRC) predisposition has been estimated to be around 35% by twin studies. Although ~ 5% of this proportion may be explained by high-penetrance mutations, and an additional 7% is thought to be due to the presence of a combination of some of the already-described 16 susceptibility SNPs, there is still a significant fraction of CRC susceptibility that remains unexplained.

On the other hand, there is also considerable variation in the way CRC patients respond to chemotherapy. Besides, the fact that most drugs used in CRC treatment have narrow therapeutic ranges results in the frequent development of adverse drug reactions (ADRs). Hence, the identification of the genetic variation modulating this outcome would be most helpful in both the individualisation of the treatment and the reduction of health costs.

1. The first aim of this study has thus been the screening for new susceptibility genetic variants in CRC. This objective is divided into two categories:

    A. The study of SNP variability. Both candidate-gene approaches and genome-wide association studies (GWAS) were used for this purpose.

    B. An evaluation on the possibility that copy-number variants (CNVs) may also be influencing CRC susceptibility.

2. The second aim of the study has been the analysis of the genetic variation underlying the differences on toxicity responses in chemotherapy-treated CRC patients.

For the first aim, we have intended to search for new variants that could explain at least a part of the missing heritability in CRC. For this purpose, we have chosen to investigate the most common sources of variability in the genome: SNPs and CNVs.

In the SNP part of the study, we have followed two different approaches: a *candidate-gene* strategy evaluating the polymorphic variation in genes with a potential functional implication in CRC carcinogenesis and a *genome-wide association study*. For the former, we have assayed in separate studies the genes present in the human syntenic regions of the 15 *Scc* (susceptibility to colorectal cancer) mouse loci (**chapter 1**), and those belonging to two pathways that have been consistently linked to CRC development: Wnt and BMP (**chapter 2**). For the latter, we have carried out a GWAS in a Spanish cohort (**chapter 3**). The advantage of this strategy against the candidate-gene one is that there is no *a priori* hypothesis on where the susceptibility loci may be located.

Regarding the CNV study, we have also performed a GWAS scan of the genomic structural variation and its potential implication in CRC neoplasia (**chapter 4**), using two different copy-number calling algorithms: Birdsuite´s *Birdseye* and QuantiSNP v2.

In the second part of this study, our purpose was to analyse the relationship between common genetic variation and the development of ADRs after chemotherapy. For this, we evaluated the correlation between two of the most common administered drugs in CRC treatment: 5-fluorouracil and oxaliplatin (FOLFOX) and the presence of ADRs by screening both SNP and CNV markers at a genome-wide level (**chapter 5**).

RESULTS

Chapter 1:

Colorectal Cancer Susceptibility Quantitative Trait Loci in Mice as a Novel Approach to Detect Low-Penetrance Variants in Humans: A Two-Stage Case-Control Study

Cancer Epidemiology, Biomarkers&Prevention (2010) Feb;*19(2):619-23.*

Cancer Epidemiology, Biomarkers & Prevention

Null Results in Brief

Colorectal Cancer Susceptibility Quantitative Trait Loci in Mice as a Novel Approach to Detect Low-Penetrance Variants in Humans: A Two-Stage Case-Control Study

Ceres Fernández-Rozadilla[1], Rosa Tarrío[1], Juan Clofent[2], Luisa de Castro[3], Alejandro Brea-Fernández[1], Xavier Bessa[4], Anna Abulí[4], Montserrat Andreu[4], Rodrigo Jover[5], Rosa Xicola[6], Xavier Llor[6], Antoni Castells[7], Sergi Castellví-Bel[7], Angel Carracedo[1], and Clara Ruiz-Ponte[1] for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association

Abstract

Thirty-five percent of colorectal cancer (CRC) susceptibility is thought to be attributable to genetics, but only a small proportion of the cases (<6%) can be explained by highly penetrant mutations. The rest of the susceptibility could be explained by a number of low-penetrance variants following a polygenic model of inheritance. Genetic modeling in rodents has been a successful tool for the unraveling of the genetic basis of diseases. The investigation of mouse quantitative trait loci led to the discovery of 15 "susceptibility to colorectal cancer" (*Scc*) loci. Thus, we aimed to analyze the human-mouse syntenic regions defined by these *Scc* loci and select human candidate genes within. Twenty-one genes were chosen and their single-nucleotide polymorphisms were tested as possible low-penetrance variants predisposing to CRC risk. Our most strongly associated single-nucleotide polymorphism, rs954353, seems to be in the 5′ region of the *CYR61* gene, which could implicate it in terms of the *cis*-regulation of the gene. *CYR61* has been proposed as a connection point among signaling pathways and a probable marker for early CRC detection. However, we could not replicate the association. Despite our negative results, we believe that our candidate gene selection strategy could be quite useful in the future determination of variants predisposing to disease. *Cancer Epidemiol Biomarkers Prev; 19(2); 619–23. ©2010 AACR.*

Introduction

Colorectal cancer (CRC) is the second most frequent neoplasm and one of the most important morbidity causes in the developed world (1). Despite the fact that 35% of CRC susceptibility could be attributable to genetics, only a small proportion of the cases (<6%) can be explained by highly penetrant mutations, suggesting that the rest of the susceptibility should exist in the form of low-penetrance variants following a polygenic model of inheritance (2).

Genetic modeling in rodents has been proved to be an important tool in the unraveling of the genetic basis of diseases. The investigation of mouse quantitative trait loci (QTL) to identify chromosomal regions harboring genetic variants that affect susceptibility successfully led to the discovery of 15 "susceptibility to colorectal cancer" (*Scc*) loci (3, 4). Because there is increasing evidence that causal genes underlying disease QTLs are conserved between rodents and humans (5), a sensible approach to identify these genes would be to map them in mice and, subsequently, investigate the role of their human homologues.

Hence, our aim is to analyze the human-mouse syntenic regions defined by these *Scc* loci and select human candidate genes to screen their single-nucleotide polymorphisms (SNP) and test them as possible low-penetrance variants predisposing to CRC risk in a two-stage case-control study.

Materials and Methods

Study Populations

Subjects on stage I were 515 CRC cases and 515 controls from EPICOLON I, a prospective, multicenter, population-based epidemiology study (6). Subjects on stage II (933 cases and 955 controls) belonged to

Authors' Affiliations: [1]Galician Public Foundation of Genomic Medicine (FPGMX), CIBERER, Genomics Medicine Group, Hospital Clínico, Santiago de Compostela, University of Santiago de Compostela, Galicia, Spain; [2]Gastroenterology Department, Hospital La Fe, Valencia, Spain; [3]Gastroenterology Department, Hospital Meixoeiro, Vigo, Galicia, Spain; [4]Gastroenterology Department, Hospital del Mar, Barcelona, Catalonia, Spain; [5]Unidad de Gastroenterología, Hospital General Universitario de Alicante, Alicante, Spain; [6]Section of Digestive Diseases and Nutrition, University of Illinois at Chicago, Chicago, Illinois; and [7]Gastroenterology Department, Hospital Clínic, CIBEREHD, IDIBAPS, Catalonia, Barcelona, Spain

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

All authors are listed in Supplementary Note.

Corresponding Author: Clara Ruiz-Ponte, Fundacion Publica Galega de Medicina Xenomica, Edificio Consultas, planta-2, Hospital Clínico, 15706 Santiago de Compostela, Galicia, Spain. Phone: 34-981-951490; Fax: 34-981-951473. E-mail: clara.ruiz.ponte@usc.es

doi: 10.1158/1055-9965.EPI-09-1175

Table 1. Description of the 15 *Scc* loci and the selected genes within the human-mouse QTL syntenic regions

QTL	Mouse chr	Human gene	Human mapping	Gene description	Gene ontology	SNPs analyzed
Scc1	2	*PTPRJ*	11p11.2	Protein tyrosine-phosphatase receptor type J	Regulation of cellular growth, differentiation and oncogenic transformation	rs10742827; rs100838801; rs10838810; rs11039519; rs1503185; rs1566734; rs2270992; rs2270993; rs4752904; rs7117386; rs7123436; rs7947811
Scc2	2	*CRB2*	9q33.2	Crumbs homolog 2	Polarized cell morphogenesis	rs10818812; rs1105223; rs1891632; rs1891638; rs33984675; rs4838051; rs7033144; rs884320
Scc3	1	*TGFB2*	1q41	Transforming growth factor β2	Suppressive effects on interleukin-2–dependent T-cell growth	rs10863396; rs1539399; rs17558745; rs1890994; rs1891467; rs2000220; rs2796821; rs4846476; rs4846479
Scc4	17	*PRKD3*	2p22-p21	Protein kinase D3	Receptor of phorbol esters: a class of tumor promoters	rs10177176; rs10460527; rs1056021; rs11124575; rs11887618; rs2300880; rs2300771; rs2300894; rs2302650; rs3770761
		MSH2	2p21	MutS homolog 2	DNA mismatch repair	rs13019654; rs17036614; rs458314; rs7607076
Scc5	18	*TNFAIP8*	5q23.1	Tumor necrosis factor α–induced protein 8	Negative mediator of apoptosis with a role in tumor progression	rs10077888; rs1045241; rs1045242; rs11064; rs17385413; rs3203922; rs32658; rs3797339; rs3797345
Scc6	11	*EGFR*	7p12	Epidermal growth factor receptor	Cell growth and differentiation control	rs1015793; rs1050171; rs1140475; rs11487218; rs11971997; rs12538489; rs12671550; rs17172446; rs17290169; rs17337023; rs2072454; rs2293347; rs3800827; rs4947492; rs4947971; rs6593205; rs6972246; rs759170; rs759171; rs7796139; rs7809394; rs88425
Scc7	3	*CYR61*	1p31-p32	Cysteine-rich 61	Promotes cell proliferation, chemotaxis, angiogenesis, and cell adhesion	rs12086058; rs12239954; rs1576424; rs3753793; rs721471; rs954353; rs9658584
Scc8	8	*TFDP1*	13q34	Transcription factor Dp-1	Regulation of the expresion of cellular promoters	rs2316121; rs6577058; rs9577286
		CDC16	13q34	Cell division cycle 16 homolog	Ubiquitin ligase with role in cell cycle control	rs3211416; rs7318644; rs7994151; rs8002514; rs9590408; rs9590409

(*Continued on the following page*)

Table 1. Description of the 15 *Scc* loci and the selected genes within the human-mouse QTL syntenic regions (Cont'd)

QTL	Mouse chr	Human gene	Human mapping	Gene description	Gene ontology	SNPs analyzed
Scc9	10	*MDM2*	12q14.3-q15	Transformed 3T3 cell double minute 2	p53 inhibitor	rs1470383; rs1795481; rs769412
		LGR5	12q22-q23	Leucine-rich repeat–containing G-protein–coupled receptor 5	Overexpressed in human colon tumors	rs10748178; rs10784923; rs11178798; rs11178832; rs11178845; rs1148985; rs12422259; rs12829521; rs17109799; rs17109924; rs17109926; rs1880892; rs3803033; rs389150; rs3923863; rs7298504; rs941197
Scc11	4	*HEYL*	1p34.3	Hairy/enhancer-of-split related with YRPW motif-like	Downstream effector of Notch signaling that networks together with Wnt	rs1180320; rs4660892; rs784622
		MYCL1	1p34.2	V-myc myelocytomatosis viral oncogene homolog 1	Loss of heterozygosity at MYCL1 is a marker for poor prognosis in CRC	rs3117088; rs3134614; rs3134615
Scc12	7	*DMBT1*	10q25.3-26	Deleted in malignant brain tumors 1	Role in the interaction of tumor cells and the immune system	rs1051715; rs2981783; rs3013236
Scc13	6	*TRAF2*	9q34	TNF receptor–associated factor 2	Regulates TNF-induced apoptosis	rs10870140; rs2784078; rs2784075; rs908831
Scc14	10	*LATS1*	6q24-q25.1	Large tumor suppressor homolog 1 (*Drosophila*)	Maintenance of ploidy and tumor supressor activity through regulation of p53	rs3798761; rs3924871
		VIP	6q25	Vasoactive intestinal peptide	Proangiogenic factor	rs12212849; rs3823082; rs637572; rs671330; rs680314; rs688136
Scc15	11	*LLGL1*	17p11.2	Lethal giant larvae homolog 1 (*Drosophila*)	Reduced expression related to progression of colon cancer; similar to a tumor supressor in *Drosophila*	rs11869582; rs2245430; rs2245737; rs2290505; rs2746027; rs8821
Ccs1	12	*FOS*	14q24.3	v-fos FBJ murine osteosarcoma viral oncogene homolog	Signal transduction protein implicated in cell proliferation and differentiation	rs1046117; rs1569328; rs3742769; rs7101
		JDP2	14q24.3	Jun dimerization protein 2	Mediator in UV-induced apoptosis, cell differentiation, tumorigenesis, and angiogenesis	rs10057; rs10873278; rs1474503; rs175644; rs4899566; rs84044

NOTE: For some of the *Scc* loci, more than one gene was selected because of their possible functional implications.

EPICOLON II, an extension of EPICOLON I. Cases and controls were matched for sex and age. All samples were obtained with informed consent reviewed by the ethical board of the corresponding hospital.

Candidate Gene Selection

QTLs were defined by their flanking markers by revision of the author's data and the MGI (7). Genes within each human-mouse syntenic region showing enriched

expression in primary affected tissues in mice were selected with ExQuest (8). Finally, 21 human genes were chosen from the 15 *Scc* (Table 1; ref. 9).

SNP Selection and Genotyping

One hundred forty-seven SNPs were selected from the 21 genes with PupaSuite (10), FESD (11), dbSNP (12), and HapMap Phase II (genome build 36; ref. 13). SNPs with unadjusted *P* values <0.01 were replicated in an independent case-control series. Genotyping was done in the SNPlex (Applied Biosystems), MassARRAY (Sequenom, Inc.), and TaqMan (Applied Biosystems) platforms at the Santiago de Compostela node of the Spanish Genotyping Center.

Statistical Analyses

Quality control was assessed with the Genotyping Data Filter (14) and Structure v2.2 (15). Genotypic distributions in controls followed Hardy-Weinberg equilibrium, and there was no sign of underlying population stratification. Association was evaluated for every single SNP and all possible haplotypes in each gene with Haploview v4.0 (16) and Unphased (17). Permutation tests and Bonferroni were used for multiple-testing corrections. Odds ratio (OR) and 95% confidence intervals were calculated with PLINK v1.03 (18). Descriptive information and association data for all the SNPs that passed quality control are shown in Supplementary Table S1.

Results

Allelic association tests revealed only one significant SNP after multiple-testing correction: rs12086058, lying in an intergenic region 6.4 kb upstream the *CYR61* gene (1p31-p22). The OR value for this SNP showed a protective effect of the minor allele (Table 2). Haplotype analysis and comparisons between sporadic and familial groups did not yield any significant associations (data not shown).

Linkage disequilibrium analysis in the *CYR61* region showed rs12086058 to be in high correlation with rs954353 ($r^2 = 1$). This SNP was located 1.8 kb upstream *CYR61*, which suggested a possible implication in the *cis*-regulation of the gene. Genotyping of rs954353 yielded a better association value than rs12086058 (2×10^{-4}). OR also showed a protective effect of the minor allele (Table 2).

To verify the results, SNPs with nominal $P < 0.01$ (rs12086058, rs954353, and rs10077888) were further replicated on an independent sample. Nevertheless, none of the associations could be replicated (Table 2).

Discussion

Our study combines the advances in CRC genetics in animal models with the investigation of the variation underlying the disease in humans. We selected 21 genes identified from syntenic regions defined by mouse QTLs to screen their SNP variability in a two-stage case-control association study. However, we did not find any replicable association. Our study had enough power to detect OR ≥1.3, assuming allelic association and $\alpha = 0.05$ (19). Results in stage I were therefore simply due to chance or to type I error.

Nevertheless, our most strongly associated SNP, rs954353, seems to be in the 5′ region of the *CYR61* gene, which could still implicate it in terms of *cis*-regulation. We analyzed the region harboring rs954353 and found it to be lying very close to two transcription factor binding site sequences. The direct sequencing of these failed to find any common variants within the consensus target that could explain the association signal found in stage I. However, we did find a 6-bp insertion polymorphism 38 bp upstream the first transcription factor binding site. This variant showed significant differences in frequencies between cases and controls ($P = 0.0236$), although no further implications could be stated about its relationship with CRC susceptibility (data not shown).

CYR61 has been proposed as a connection point among signaling pathways and a probable marker for early CRC detection (20). Besides, it has been extensively implicated in carcinogenesis-related events such as angiogenesis (21), tissue invasion (22), cell migration, and metastasis (23), although no association studies have been published thus far that analyze its relationship with CRC.

Despite our negative results, we believe that our candidate gene selection, through the identification of genes or regions conferring susceptibility to other species, could be quite useful in the future determination of variants predisposing to disease. Our QTLs analyses proved to be very helpful as a starting point in the search for candidate genes affecting CRC susceptibility because all the genes identified were somehow related to carcinogenetic events.

Table 2. Association analyses for the three SNPs selected for replication on stage II

SNP_ID	Gene	Relevance	Alleles	Observed MAF	OR (95% CI)	$\chi^{2\ 1df}$ *P*	Stage I permutations *P*	Bonferroni *P*	Stage II $\chi^{2\ 1df}$ *P*
rs12086058	*CYR61*	5′UTR	A/G	0.428	0.71 (0.59-0.86)	0.0005	0.0326	0.0405	0.4099
rs954353	*CYR61*	5′UTR	A/G	0.434	0.70 (0.59-0.84)	0.0002	0.0246	0.027	0.3917
rs10077888	*TNFAIP8*	Intronic	C/G	0.302	0.75 (0.61-0.92)	0.0019	0.2058	0.2565	0.8188

Abbreviations: MAF, minor allele frequency; 95% CI, 95% confidence interval; UTR, untranslated region.

In fact, although this approach has not been successful thus far for CRC, it positively identified a haplotype in *PTPRJ* as a breast cancer genetic susceptibility low-penetrance allele (24). Hence, we encourage future efforts in this field and believe that the relationship between *CYR61* and CRC should be studied in other populations to fully discard a putative genetic association.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

Acknowledgments

We thank all the patients that participated in this study, who were recruited in 25 Spanish hospitals as part of the EPICOLON project. S. Castellví-Bel is supported by a contract from the Fondo de Investigación Sanitaria (CP 03-0070). C. Fernández-Rozadilla has obtained a FPU Fellowship from the Ministerio de Educacion; CIBERER and CIBEREHD are funded by Instituto de Salud Carlos III. We thank Maria Magdalena-Castro, Olga Lortes, and Eva Fernández for their excellent technical assistance. M. Magdalena-Castro and E. Fernández are supported by Isabel Barreto's program from Xunta de Galicia, and O. Lortes by a contract from the CIBERER.

Grant Support

Fondo de Investigación Sanitaria/FEDER (06/1384, 08/0024, 08/1276), Fundación Mutua Madrileña (C. Ruiz-Ponte and S. Castellví-Bel), Ministerio de Educación y Ciencia (SAF 07-64873), Asociación Española contra el Cáncer, Fundación Olga Torres (S. Castellví-Bel), Acción en Cáncer (Instituto de Salud Carlos III), and Xunta de Galicia (RHI07/04).

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received 11/20/09; revised 12/3/09; accepted 12/7/09; published online 2/8/10.

References

1. Ferlay J, Bray F, Pisani P, Parkin DM. GLOBOCAN 2002: cancer incidence, mortality and prevalence worldwide. IARC Cancer Base No. 5. Version 2.0. IARC Press: Lyon 2004.
2. Houlston RS, Peto J. The search for low-penetrance cancer susceptibility alleles. Oncogene 2004;23:6471–6.
3. Demant P. Cancer susceptibility in the mouse: genetics, biology and implications for human cancer. Nat Rev Genet 2003;4:721–34.
4. Ruivenkamp CA, van Wezel T, Zanon C, et al. *Ptprj* is a candidate for the mouse colon-cancer susceptibility locus Scc1 and is frequently deleted in human cancers. Nat Genet 2002;31:295–300.
5. Peters LL, Robledo RF, Bult CJ, Churchill GA, Paigen BJ, Svenson KL. The mouse as a model for human biology: a resource guide for complex trait analysis. Nat Rev Genet 2007;8:58–69.
6. Piñol V, Castells A, Andreu A, et al. Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis colorectal cancer. JAMA 2005;293:1986–94.
7. Mouse Genomic Informatics. http://www.informatics.jax.org/.
8. ExQuest. http://dev.thep.lu.se/proteios/wiki/ExQuest.
9. The Gene Ontology. http://www.geneontology.org/.
10. Conde L, Vaquerizas JM, Dopazo H, et al. PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res 2006;34:W621–5.
11. Kang HJ, Choi KO, Kim BD, Kim S, Kim YJ. FESD: a Functional Element SNPs Database in human. Nucleic Acids Res 2005;33:D518–22.
12. dbSNP. Available at: http://www.ncbi.nlm.nih.gov/projects/SNP/.
13. The International HapMap Consortium. The International HapMap Project. Nature 2003;426:789–96.
14. Genotyping Data Filter Software. Available at: http://bioinformatics.cesga.es/gdf/nav_input.php.
15. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics 2000;155:945–59.
16. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:263–5.
17. Dudbridge F. Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 2003;25:115–21.
18. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a toolset for whole-genome association and population-based linkage-analysis. Am J Hum Genet, 81. Available at: http://pngu.mgh.harvard.edu/~purcell/plink/.
19. Genetic Power Calculator. Available at: http://pngu.mgh.harvard.edu/~purcell/gpc/.
20. Hong Y, Ho KS, Eu KW, Cheah PY. A susceptibility gene set for early onset colorectal cancer that integrates diverse signaling pathways: implication for tumorigenesis. Clin Cancer Res 2007;13:1107–14.
21. Gashaw I, Stiller S, Boing C, Kimmig R, Winterhager E. Premenstrual regulation of the pro-angiogenic factor *CYR61* in human endometrium. Endocrinology 2008;149:2261–9.
22. Monnier Y, Farmer P, Bieler G, et al. CYR61 and $\alpha_V\beta_5$ integrin cooperate to promote invasion and metastasis of tumors growing in preirradiated stroma. Cancer Res 2008;68:7323–31.
23. Sun ZJ, Wang Y, Cai Z, Chen PP, Tong XJ, Xie D. Involvement of CYR61 in growth, migration, and metastasis of prostate cancer cells. Br J Cancer 2008;99:1656–67.
24. Lesueur F, Pharoah PD, Laing S, et al. Allelic association of the human homologue of the mouse modifier *Ptprj* with breast cancer. Hum Mol Genet 2005;14:2349–56.

Chapter 2:
Single Nucleotide Polymorphisms in the Wnt and BMP Pathways and Colorectal Cancer Risk in a Spanish Cohort

Plos One (2010) Sep 9;5(9). pii: e12673.

OPEN ACCESS Freely available online

PLoS one

Single Nucleotide Polymorphisms in the Wnt and BMP Pathways and Colorectal Cancer Risk in a Spanish Cohort

Ceres Fernández-Rozadilla[1], Luisa de Castro[2], Juan Clofent[3], Alejandro Brea-Fernández[1], Xavier Bessa[4], Anna Abulí[4], Montserrat Andreu[4], Rodrigo Jover[5], Rosa Xicola[6], Xavier Llor[6], Antoni Castells[7], Sergi Castellví-Bel[7], Angel Carracedo[1], Clara Ruiz-Ponte[1]*, for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association[¶]

1 Galician Public Foundation of Genomic Medicine (FPGMX), Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER), Genomics Medicine Group, Hospital Clínico, Santiago de Compostela, University of Santiago de Compostela, Galicia, Spain, **2** Gastroenterology Department, Hospital Meixoeiro, Vigo, Galicia, Spain, **3** Gastroenterology Department, Hospital La Fe, Valencia, Spain, **4** Gastroenterology Department, Hospital del Mar, Institut Municipal d'Investigació Médica (IMIM), Pompeu Fabra University, Barcelona, Catalonia, Spain, **5** Unidad de Gastroenterología, Hospital General Universitario de Alicante, Alicante, Spain, **6** Section of Digestive Diseases and Nutrition, University of Illinois at Chicago, Chicago, Illinois, United States of America, **7** Department of Gastroenterology, Hospital Clínic, CIBERehd, IDIBAPS, University of Barcelona, Barcelona, Catalonia, Spain

Abstract

Background: Colorectal cancer (CRC) is considered a complex disease, and thus the majority of the genetic susceptibility is thought to lie in the form of low-penetrance variants following a polygenic model of inheritance. Candidate-gene studies have so far been one of the basic approaches taken to identify these susceptibility variants. The consistent involvement of some signaling routes in carcinogenesis provided support for pathway-based studies as a natural strategy to select genes that could potentially harbour new susceptibility loci.

Methodology/Principal Findings: We selected two main carcinogenesis-related pathways: Wnt and BMP, in order to screen the implicated genes for new risk variants. We then conducted a case-control association study in 933 CRC cases and 969 controls based on coding and regulatory SNPs. We also included rs4444235 and rs9929218, which did not fulfill our selection criteria but belonged to two genes in the BMP pathway and had consistently been linked to CRC in previous studies. Neither allelic, nor genotypic or haplotypic analyses showed any signs of association between the 37 screened variants and CRC risk. Adjustments for sex and age, and stratified analysis between sporadic and control groups did not yield any positive results either.

Conclusions/Significance: Despite the relevance of both pathways in the pathogenesis of the disease, and the fact that this is indeed the first study that considers these pathways as a candidate-gene selection approach, our study does not present any evidence of the presence of low-penetrance variants for the selected markers in any of the considered genes in our cohort.

Citation: Fernández-Rozadilla C, de Castro L, Clofent J, Brea-Fernández A, Bessa X, et al. (2010) Single Nucleotide Polymorphisms in the Wnt and BMP Pathways and Colorectal Cancer Risk in a Spanish Cohort. PLoS ONE 5(9): e12673. doi:10.1371/journal.pone.0012673

Editor: Chun-Ming Wong, University of Hong Kong, Hong Kong

Received May 20, 2010; **Accepted** August 6, 2010; **Published** September 9, 2010

Funding: This work was supported by grants from the Fondo de Investigacion Sanitaria/FEDER (05/2031, 08/0024, 08/1276, PS09/02368), Xunta de Galicia (PGIDIT07PXIB9101209PR, Ministerio de Ciencia e Innovacion (SAF 07-64873), Asociacion Espanola contra el Cancer (Fundacion Cientifica y Junta de Barcelona), and Fundacion de Investigacion Medica Mutua Madrilena. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: clara.ruiz.ponte@usc.es

¶ Members of the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association are listed in Note S1

Introduction

Colorectal cancer (CRC) is one of the main forms of cancer, being the second most frequent neoplasm in both sexes and one of the most important morbidity causes in the western world [1]. The genetic contribution to CRC has been estimated to be around 35% by extensive twin studies [2]. However, highly penetrant variants, that cause mendelian predisposition syndromes, account only for, at most, 5% of the disease cases [3]. The remaining genetic susceptibility is thought to follow a polygenic model, with an interplay of multiple low-penetrance allelic variants appearing in high frequency in the general population, and each conferring a modest effect on disease risk [4,5].

Candidate-gene studies have been one of the most commonly used tools in the screening for new variants affecting CRC risk. Gene selection in these studies is mainly based on the functional implications of a possible association, and thus genes selected have either been chosen because of the previous presence of other high/low risk alleles [6], or their participation in a pathway implicated in the pathogenesis of the disease [7]. Candidate-gene studies can be performed by either direct approaches, where the variants genotyped are presumed to be the underlying cause of the disease

because of their location (variants in exonic or regulatory regions), or by indirect approaches, where tag SNPs take advantage of the linkage disequilibrium properties of the human genome to try and screen the most of the variability in a given gene.

This latter approach has also allowed, together with the development of high-throughput technologies, the implementation of new hypothesis-free approaches (in opposition with hypothesis-based candidate-gene approaches), covering the majority of the genome (genome-wide association studies or GWAS). This implementation has successfully led to the identification of some new susceptibility loci [8–14], including rs4444235 and rs9929218, that fall within reach of two genes belonging to the BMP pathway. Nevertheless, these have been found to predict only a small proportion of the disease susceptibility, with the remaining yet to be discovered [15].

We hence aimed to find such susceptibility variants through a candidate-gene approach screening a selected number of variants within two cellular pathways that have consistently been linked to CRC tumorogenesis: the Wnt and the BMP signaling pathways [16,17].

The Wnt pathway contains genes that have for long been known to be responsible of some hereditary CRC syndromes, such as *APC* and familial adenomatous polyposis [18]. Moreover, somatic alterations in *APC* are found in almost 80% of the sporadic colorectal cancers, and Wnt signaling activation is involved in the best part of sporadic colorectal carcinomas [19]. On the other hand, the BMP pathway acts as positive regulator of some of the Wnt proteins [17], and the tumor suppressive role of this signaling pathway in the pathogenesis of CRC and other cancers is well established [20,21]. Besides, mutations in two of its genes, *SMAD4* and *BMPR1A*, are responsible for juvenile polyposis syndrome, another hereditary CRC condition [22]. Considering all this information, we thought it would be interesting to screen some of the genetic variability within these pathways for any evidence of new CRC related variants that could explain at least part of the missing heritability. Our approach was mainly functional, for only SNPs within exonic or *cis*-regulatory sequences (5′ and 3′ unstranslated regions) were selected to analyse their relationship with CRC susceptibility.

Results and Discussion

Following our pathway-based candidate-gene selection method, we performed our study in a total of 45 SNPs that were in either exonic or regulatory regions, in an overall of 21 genes from both the Wnt and BMP pathways. Details of SNP features and association values for the 37 SNPs that successfully passed quality control criteria are shown on Table 1. None of the screened SNPs were significantly associated with an altered risk of CRC, considering odds-ratios and related p values for allelic and genotypic tests (trend, dominant and recessive). Logistic regression for age and sex adjustment was performed, although it did not improve p value results. Haplotype analysis results were consistent in both Unphased and Haploview, and did not show any signs of positive associations either for any of the 8 genes for which this analysis was performed (*AXIN1*, *HDAC9*, *BMP4*, *DACT1*, *CDH3*, *CDH1*, *BTRC*, and *APC)*, (Figure S1). Stratification analysis comparing sporadic and familial cases was also implemented, but it did not provide any evidence of differences in susceptibilities between the groups that could be a sign of any specific associations within either of the groups (Table 2).

Thus, our strategy has not managed to detect any new susceptibility loci for CRC risk.

Pathway-based expectations have proved to be quite discouraging in the literature as well, for strong candidate pathways, such as DNA-repair ones, surprisingly failed too in identifying any new risk variants [7,23–24]. In addition to this, most of the genetic variants that have been found to be associated with disease are located in intergenic regions, with potential functions that are yet unknown.

Still, in light of the recent discoveries that followed up the analysis of genome-wide data, both Wnt and BMP have earned a renewed fame. The susceptibility locus found on 8q24 (rs6983267) has been linked to an enhanced Wnt signaling through its interaction with TCF4 [25,26], and a meta-analysis conducted on a series of GWAS data succeeded in associating two variants in the *BMP4* and *CDH1* gene regions with the disease (rs4444235 and rs9929218, respectively)[8].

Even though this is actually the first association study that considers the pathways as a whole for gene selection, some of the genes included in our analysis (i.e *APC, CCND1*, *CDH1* and *TCF7*) had already been screened for risk alleles [6,27–30]. It is quite remarkable that there has been a growing debate over some of these loci, specially the p.V1822D variant in *APC* (rs459552). This missense change is widely documented in the literature, with some studies defending it as neutral (this study and others)[31], and some conferring its minor allele a protective effect [6,28]. Lack of appropriate study power, resultant from insufficient number of samples has been a major problem in many of these studies and thus most of them have not provided very convincing results [32].

Although our study had over 80% power to detect OR as low as 1.21 with minor allele frequencies of 0.30 (57% of our SNPs), and 1.24 for MAFs down to 0.2 (78% of the SNPs), assuming a log-additive model and $\alpha = 0.05$, we were unable to detect any positive associations suggesting the presence of any new CRC susceptibility variants. Nevertheless, it is quite remarkable that, albeit our failure to replicate the associations for the *BMP4* and *CDH1* SNPs, this is the first study that investigates any of the so-called 10 new GWAS-discovered susceptibility loci in a Southern-European population.

Despite our negative results, we must consider that we did not whatsoever comprehensively cover all possible low-penetrance variants within the selected genes. This is mainly due to the fact that our strategy was purely functional, selecting the variants that were *a priori* good candidates to be directly associated with the disease. This indeed may constitute a limitation in the study, for most of the genetic variation within the loci was not investigated. Thus, we believe further efforts should be made to screen a wider variety of loci within these pathways, specially considering the previous positive associations described so far for both Wnt and BMP-related genes.

Pondering the potential odds ratios of the variants described so far (1.11, CI 1.08–1.15 and 0.91, CI 0.89–0.94 for rs4444235 and rs9929218, respectively), we assume larger cohorts may be required to detect such subtle effects. On the other hand, when considering candidate-gene approaches, it would also be useful to meta-analyse previous studies and pull the information across of them altogether in the search of evidences of potential new pathways linked to the pathogenesis of the disease.

Materials and Methods

Study populations

Subjects were 933 CRC patients and 969 controls that belonged to the EPICOLON project, a prospective, multicentre, population-based epidemiology survey studying the incidence and features of familial and sporadic CRC in the Spanish population

Table 1. Description of the 37 SNPs that passed quality control criteria and their associated p values.

Gene	SNP ID	SNP type	Amino acid change	Allele	MAF cases	MAF controls	GT counts cases	GT counts controls	p-value	OR (95% CI)
ADAR	rs2229857	Missense	K384R	**A**/G	0.3306	0.3201	99/360/385	88/347/382	0.512	1.05 (0.91–1.22)
APC	rs2229992	Synonymous	Y486Y	C/**T**	0.3981	0.4125	145/382/317	141/392/284	0.3728	0.94 (0.82–1.08)
APC	rs351771	Synonymous	A545A	**C**/T	0.3817	0.375	124/397/324	125/416/347	0.7978	1.02 (0.89–1.18)
APC	rs41115	Synonymous	T1493T	**C**/T	0.3796	0.3761	126/385/328	127/414/347	0.8595	1.00 (0.88–1.16)
APC	rs42427	Synonymous	G1678G	A/**G**	0.3741	0.3713	124/382/336	116/365/323	0.9252	1.01 (0.88–1.17)
APC	rs459552	Missense	V1822D	**A**/T	0.2302	0.2134	48/293/504	41/297/550	0.2197	1.11 (0.94–1.30)
APC	rs465899	Synonymous	P1960P	**C**/T	0.3828	0.3743	126/395/324	125/414/348	0.7107	1.03 (0.90–1.19)
APC	rs866006	Synonymous	S1756S	**A**/C	0.3775	0.3756	123/370/323	124/401/339	0.925	1.00 (0.87–1.19)
AXIN1	rs1805105	Synonymous	D254D	C/**T**	0.3918	0.4096	136/387/318	164/397/324	0.2692	0.93 (0.81–1.07)
AXIN1	rs214250	Synonymous	S428S	C/**T**	0.2206	0.2028	32/307/502	34/265/522	0.2138	1.12 (0.94–1.32)
AXIN1	rs214252	Synonymous	A609A	A/**G**	0.2207	0.2005	32/305/499	34/258/521	0.1403	1.13 (0.96–1.34)
AXIN1	rs400037	Missense	R388Q	C/**T**	0.1826	0.1829	27/244/545	39/234/580	0.8972	1.04 (0.87–1.24)
AXIN2	rs2240308	Missense	P50S	**A**/G	0.4502	0.4219	168/423/252	152/442/290	0.1031	1.11 (0.97–1.27)
BMP4	rs17563	Missense	V152A	C/**T**	0.4946	0.4855	211/407/220	208/420/233	0.5498	1.07 (0.93–1.23)
BMP4	rs4444235	–	–	C/**T**	0.4563	0.4557	168/436/242	196/411/274	0.9343	0.99 (0.86–1.14)*
BTRC	rs17767748	Synonymous	I229I	C/**T**	0.05516	0.056	3/86/745	4/91/789	0.9324	1.00 (0.74–1.36)
BTRC	rs4151060	Missense	A543S	G/**T**	0.04793	0.04904	4/73/768	2/83/802	0.6997	0.96 (0.70–1.32)
CCND1	rs603965	Synonymous	P241P	**A**/G	0.4969	0.4822	204/406/209	206/430/237	0.4164	1.06 (0.93–1.22)
CDH1	rs1801552	Synonymous	A692A	C/**T**	0.3547	0.3781	105/371/343	126/365/325	0.1834	0.92 (0.81–1.07)
CDH1	rs9929218	Intronic	–	**A**/G	0.2811	0.2873	65/345/435	83/342/459	0.5486	0.97 (0.83–1.13)*
CDH3	rs1126933	Missense	Q563H	**C**/G	0.3828	0.3802	129/382/325	129/361/324	0.8369	1.02 (0.88–1.17)
CDH3	rs17715450	Synonymous	R747R	A/**C**	0.3783	0.3959	116/390/316	147/402/330	0.2792	0.93 (0.80–1.07)
CDH3	rs2274239	Synonymous	K652K	**C**/T	0.3599	0.3771	108/390/344	126/368/328	0.2863	0.93 (0.81–1.07)
CDH3	rs2296408	Synonymous	T271T	**G**/T	0.3698	0.3724	107/394/321	130/388/352	0.8768	1.00 (0.87–1.15)
CDH3	rs2296409	Synonymous	T240T	**C**/T	0.3585	0.3643	106/391/344	130/387/371	0.7962	0.98 (0.85–1.13)
CDH3	rs8049247	Synonymous	I204I	**A**/C	0.1665	0.1682	21/238/582	22/249/600	0.8683	0.97 (0.81–1.17)
DACT1	rs17832998	Missense	A464V	C/**T**	0.3468	0.3448	111/362/369	116/381/392	0.9293	1.01 (0.88–1.17)
DACT1	rs863091	Synonymous	V378V	C/**T**	0.2047	0.2033	30/283/525	41/249/524	0.932	1.01 (0.85–1.19)
HDAC9	rs1178127	Missense	P621P	A/**G**	0.21	0.2203	37/273/516	41/300/526	0.4737	0.94 (0.80–1.12)
HDAC9	rs34096894	Synonymous	L152L	C/**T**	0.01953	0.01351	0/33/812	1/22/865	0.2075	1.33 (0.78–2.27)
NLK	rs3182380	Synonymous	I498I	C/**T**	0.05142	0.05535	2/83/761	3/85/734	0.4686	0.92 (0.68–1.24)
PPARD	rs2076167	Synonymous	N163N	A/**G**	0.2956	0.294	72/355/417	78/328/417	0.9891	1.00 (0.86–1.16)
SMURF1	rs219797	Synonymous	S166S	C/**G**	0.4452	0.4712	160/428/252	210/415/261	0.1591	0.90 (0.78–1.03)
TCF7	rs30489	Missense	G256R	C/**T**	0.07683	0.07937	6/118/722	6/128/748	0.7655	0.97 (0.75–1.25)
TLE1	rs2228173	Synonymous	E118E	**A**/G	0.1183	0.1172	11/178/656	6/196/685	0.992	1.02 (0.82–1.26)
WIF1	rs7301320	Synonymous	A73A	C/**T**	0.2237	0.2219	48/265/494	47/281/517	0.9768	1.00 (0.84–1.18)
WNT2B	rs910697	Synonymous	Q390Q	A/**G**	0.4218	0.4301	154/404/286	172/419/296	0.5463	0.95 (0.83–1.09)

Minor allele is depicted in bold.

MAF. Minor Allele Frequency; OR 95% CI. Odds Ratio and 95% Confidence Interval. GT counts. Genotype counts.

**Described OR (95%CI) for rs4444235 and rs9929218 were 1.11 (1.08–1.15) and 0.91 (0.89–0.94), respectively, as taken from Houlston et al.. Nat Genet 2008.*

doi:10.1371/journal.pone.0012673.t001

[33]. Cases were selected across 11 hospitals in Spain as all patients with a *de-novo* histologically confirmed diagnosis of colorectal adenocarcinoma and who attended 11 community hospitals across Spain between November 2006 and December 2007. Patients in whom CRC developed in the context of familial adenomatous polyposis or inflammatory bowel disease, and cases where patients or family refused to participate in the study were excluded. Demographic, clinical and tumour-related characteristics of probands, as well as a detailed family history were obtained using a pre-established questionnaire, and registered in a single database. Of these, 592 (63%) were male and 341 (37%) female. Median age for cases was 73 (range 26–95), whereas mean was 71(SD±10.7). Hospital-based controls were recruited together with cases and were confirmed to have no cancer or prior history of neoplasm, and no family history of CRC. All controls were randomly selected and matched with cases for sex and age (±5 years) in a 1:1 ratio. Both cases and controls were of European ancestry and from Spain.

Table 2. Association values for stratified analysis in familial and sporadic CRC groups.

		Familial vs control		Sporadic vs control		Familial vs sporadic	
ADAR	rs2229857	0.08586	1.28 (0.97–1.68)	0.8662	1.01 (0.87–1.18)	0.1011	1.26 (0.95–1.67)
APC	rs2229992	0.6564	1.06 (0.81–1.39)	0.2732	0.92 (0.80–1.07)	0.3214	1.15 (0.87–1.51)
APC	rs351771	0.3266	1.15 (0.87–1.50)	0.8956	1.01 (0.87–1.17)	0.3659	1.14 (0.86–1.49)
APC	rs41115	0.4254	1.12 (0.85–1.47)	0.9802	1.00 (0.86–1.15)	0.4306	1.12 (0.85–1.47)
APC	rs42427	0.3978	1.13 (0.86–1.48)	0.9322	0.99 (0.86–1.15)	0.3825	1.13 (0.86–1.49)
APC	rs459552	0.05147	1.35 (1.00–1.83)	0.4821	1.06 (0.90–1.26)	0.1313	1.27 (0.93–1.72)
APC	rs465899	0.3161	1.15 (0.88–1.51)	0.8003	1.02 (0.88–1.18)	0.3885	1.13 (0.86–1.49)
APC	rs866006	0.3634	1.14 (0.86–1.49)	0.8589	0.99 (0.85–1.14)	0.3256	1.15 (0.87–1.52)
AXIN1	rs1805105	0.0674	0.77 (0.58–1.02)	0.5492	0.96 (0.83–1.10)	0.1524	0.81 (0.61–1.08)
AXIN1	rs214250	0.5041	1.12 (0.81–1.55)	0.2312	1.11 (0.93–1.32)	0.9975	1.00 (0.72–1.39)
AXIN1	rs214252	0.4511	1.13 (0.82–1.57)	0.1736	1.13 (0.95–1.34)	0.9984	1.00 (0.72–1.39)
AXIN1	rs400037	0.1971	1.25 (0.89–1.74)	0.6545	0.96 (0.80–1.15)	0.1447	1.29 (0.92–1.81)
AXIN2	rs2240308	0.7901	1.04 (0.78–1.36)	0.0733	1.14 (0.99–1.31)	0.5069	0.91 (0.69–1.20)
BMP4	rs17563	0.1037	1.25(0.95–1.64)	0.9434	1.01 (0.87–1.16)	0.1119	1.25 (0.95–1.64)
BMP4	rs4444235	0.2311	0.85 (0.65–1.11)	0.6689	1.03 (0.90–1.19)	0.1486	0.82 (0.62–1.08)
BTRC	rs17767748	0.7285	1.10 (0.63–1.93)	0.813	0.96 (0.71–1.31)	0.6361	1.15 (0.65–2.03)
BTRC	rs4151060	0.1176	1.52 (0.90–2.57)	0.4741	0.89 (0.64–1.24)	0.04729	1.72 (1.00–2.96)
CCND1	rs603965	0.335	0.87 (0.66–1.15)	0.2045	1.10 (0.95–1.26)	0.1203	0.80 (0.61–1.06)
CDH1	rs1801552	0.6563	1.07 (0.80–1.41)	0.08919	0.88 (0.76–1.02)	0.1812	1.21 (0.91–1.61)
CDH1	rs9929218	0.8686	0.98 (0.73–1.31)	0.6861	0.97 (0.83–1.13)	0.926	1.01 (0.75–1.37)
CDH3	rs1126933	0.1283	1.23 (0.94–1.62)	0.7438	0.98 (0.84–1.13)	0.09059	1.27 (0.96–1.67
CDH3	rs17715450	0.2767	0.86 (0.65–1.13)	0.4126	0.94 (0.82–1.09)	0.5064	0.91 (0.68–1.21)
CDH3	rs2274239	0.1972	0.83 (0.63–1.10)	0.4589	0.95 (0.82–1.10)	0.3649	0.88 (0.66–1.17)
CDH3	rs2296408	0.4447	0.90 (0.68–1.19)	0.9386	1.01 (0.87–1.16)	0.4216	0.89 (0.67–1.18)
CDH3	rs2296409	0.1256	0.80 (0.60–1.07)	0.9158	1.01 (0.87–1.17)	0.1138	0.79 (0.59–1.06)
CDH3	rs8049247	0.9636	1.01 (0.71–1.44)	0.867	0.98 (0.82–1.19)	0.9021	1.02 (0.71–1.47)
DACT1	rs17832998	0.9185	0.99 (0.74–1.31)	0.8619	1.01 (0.88–1.17)	0.8392	0.97 (0.73–1.29)
DACT1	rs863091	0.5683	0.90 (0.64–1.28)	0.7737	1.03 (0.86–1.22)	0.4595	0.88 (0.62–1.24)
Gene	SNP ID	p-value	OR (CI 95%)	p-value	OR (CI 95%)	p-value	OR (CI 95%)
HDAC9	rs1178127	0.8693	1.03 (0.74–1.42)	0.3847	0.93 (0.78–1.10)	0.5511	1.11 (0.80–1.54)
HDAC9	rs34096894	0.8555	0.89 (0.27–2.99)	0.1093	1.55 (0.90–2.67)	0.3638	0.58 (0.18–1.91)
NLK	rs3182380	0.4747	0.79 (0.42–1.50)	0.7387	0.95 (0.69–1.30)	0.5917	0.84 (0.44–1.60
PPARD	rs2076167	0.1051	0.77 (0.57–1.06)	0.5291	1.05 (0.90–1.23)	0.06342	0.74 (0.55–1.02)
SMURF1	rs219797	0.9123	0.99 (0.75–1.29)	0.09224	0.89 (0.77–1.02)	0.4764	1.10 (0.84–1.45)
TCF7	rs30489	0.1722	1.36 (0.87–2.11)	0.4351	0.90 (0.69–1.17)	0.07095	1.51 (0.96–2.38)
TLE	rs2228173	0.4715	1.16 (0.78–1.71)	0.8995	0.99 (0.79–1.23)	0.4626	1.16 (0.78–1.73)
WIF1	rs7301320	0.2681	1.20 (0.87–1.64)	0.8226	0.98 (0.83–1.16)	0.2418	1.21 (0.88–1.67)
WNT2B	rs910697	0.4228	0.90 (0.68–1.17)	0.7713	0.98 (0.85–1.13)	0.5418	0.92 (0.70–1.21)

MAF. Minor Allele Frequency; OR 95% CI. Odds Ratio and 95% Confidence Interval.
doi:10.1371/journal.pone.0012673.t002

Ethics statement

The study was approved by the ''Comité Ético de Investigación Clínica de Galicia'', and each of the institutional review boards of the hospitals where samples were collected (''Ethics Committee of the Hospital Clínic-Barcelona'', ''Ethics Committee of the Hospital del Mar-Barcelona'', ''Ethics Committee of the Hospital German Trias i Pujol-Barcelona'', ''Ethics Committee of the Hospital Sant Pau-Barcelona,'' ''Ethics Committee of the Hospital Universitari Arnau de Vilanova-Lleida'', ''Ethics Committee of the Hospital General-Alicante'', ''Ethics Committee of the Hospital de Donosti'', ''Ethics Committee of the Hospital General de Asturias-Oviedo'', ''Ethics Committee of the Hospital Clinico-Zaragoza'', ''Ethics Committee of the Hospital de Calahorra-La Rioja'', ''Ethics Committee of the Hospital Meixoeiro-Vigo''). All samples were obtained with written informed consent reviewed by the ethical board of the corresponding hospital.

DNA extraction

DNA was obtained from frozen peripheral blood; extraction was performed in a CHEMAGEN robot (Chemagen Biopolymer-

Table 3. Description of all genes selected from both pathways and SNPs screened within each of them.

Gene Name	Function	pathway/genes modulated by BMP signalling	SNPs selected
***ADAR*, Adenosine deaminase, RNA- specific**	Converts multiple adenosines to inosines and creates I/U mismatched base pairs in double-helical RNA	Wnt signalling[36]	rs2229857
***APC*, Adenomatous Polyposis Coli**	B-catenin degradation	Wnt signalling[36]	rs2229992,rs351771,rs4115, rs42427rs459552,rs465899,rs86006
***AXIN1*, Axin 1**	B-catenin regulation	Wnt signalling[36]	rs1048786,rs1805105,rs214250, rs214252,rs400037,rs419949
***BTRC*, Beta-transducin repeat containing**	B-catenin ubiquitination	Wnt signalling[36]	rs17767748,rs415060
***CCND1*, Cyclin D1**	Cell cycle control	Wnt signalling[36]	rs603965
CSNK1A1, Casein kinase 1, alpha 1	B-catenin fosforilation	Wnt signalling[36]	NA
CSNK2A1, Casein kinase 2, alpha 1	B-catenin fosforilation	Wnt signalling[36]	NA
CTBP1, C-terminal binding protein 1	Transcriptional repressor in cellular proliferation	Wnt signalling[36]	NA
CTNNB1, Catenin (cadherin-associated protein), beta 1	Cell adhesion and signal transduction	Wnt signalling[36]	NA
EIF4E, Eukaryotic translation initiation factor 4E	Translation initiation factor	Wnt signalling[36]	NA
ELAC1, ElaC homolog 1 (E. coli)	Zinc phosphodiesterase	Wnt signalling[36]	NA
FRAT1, Frequently rearranged in advanced T-cell lymphomas	B-catenin stabilization	Wnt signalling[36]	NA
FZD1, Frizzled homolog 1 (Drosophila)	Receptor for Wnt proteins	Wnt signalling[36]	NA
***GSK3B*, Glycogen synthase kinase 3 beta**	B-catenin fosforilation	Wnt signalling[36]	rs34002644
***HDAC9*, Histone deacetylase 9**	Transcriptional regulation, cell cycle	Wnt signalling[36]	rs1178127,rs34096894
***HNF4A*, Hepatocyte nuclear factor 4, alpha**	Transcriptionally controlled transcription factor	Wnt signalling[36]	rs35078168
MAP3K7, Mitogen-activated protein kinase kinase kinase 7	Signaling transduction induced by BMP	Wnt signalling[36]	NA
MYC, v-myc myelocytomatosis viral oncogene homolog (avian)	Regulation of gene transcription	Wnt signalling[36]	NA
***NLK*, Nemo-like kinase**	Negatively regulation wnt pathway	Wnt signalling[36]	rs3182380
***PPARD*, Peroxisome proliferator-activated receptor delta**	Ligand-activated transcription factor.	Wnt signalling[36]	rs2076167
PPP2R4, Protein phosphatase 2A activator, regulatory subunit 4	Folding of proteins	Wnt signalling[36]	NA
***TLE1*, Transducin-like enhancer of split 1 (E(sp1) homolog, Drosophila)**	Transcriptional corepressor	Wnt signalling[36]	rs2228173,rs8782
***WIF1*, Wnt inhibitory factor 1**	Inhibition of the WNT activities	Wnt signalling[36]	rs1026024,rs7301320
WNT1, Wingless-type MMTV integration site family, member 1	Ligand for members of the frizzled family	Wnt signalling[36]	NA
***BMP4*, Bone morphogenetic protein 4**	Induces cartilage and bone formation.	BMP signalling[17]	rs17563
BMPR1B, Bone morphogenetic protein receptor, type IB	Transmembrane serine/threonine	BMP signalling[17]	NA
SMAD1, SMAD family member 1	Signal transduction	BMP signalling[17]	NA
***SMAD4*, SMAD family member 4**	Signal transduction	BMP signalling[17]	rs75667697
SMAD5, SMAD family member 5	Signal transduction	BMP signalling[17]	NA
***SMURF1*, SMAD specific E3 ubiquitin protein ligase 1**	Ubiquitination and degradation of SMAD proteins	BMP signalling[17]	rs219797
***AXIN2*, Axin 2**	B-catenin regulation	Wnt signalling, BMP induced genes[34]	rs2240308
***CDH1*, Cadherin 1, type 1, E-cadherin**	B-catenin regulation	Wnt signalling, BMP induced genes[34]	rs1801552
***CDH3*, Cadherin 3, type 1, P-cadherin (placental)**	B-catenin regulation	Wnt signalling, BMP induced genes[34]	rs1126933,rs17715450,rs2274239, rs2296408,rs2296409,rs8049247
DAB2, Disabled homolog 2, mitogen-responsive phosphoprotein	B-catenin regulation	Wnt signalling, BMP induced genes[34]	NA

Table 3. Cont.

Gene Name	Function	pathway/genes modulated by BMP signalling	SNPs selected
***DACT1*, Dapper antagonist of beta-catenin, homolog 1 (Xenopus laevis)**	Disheveled inhibitor	Wnt signalling, BMP induced genes[34]	rs17832998,rs698025,rs863091
KIFAP3, Kinesin-associated protein 3	Interacts with apc	Wnt signalling, BMP induced genes[34]	NA
LEF1, Lymphoid enhancer-binding factor 1	Transcriptional activator of Wnt signaling	Wnt signalling, BMP induced genes[34]	NA
***TCF7*, Transcription factor 7 (T-cell specific, HMG-box)**	transcriptional repressor of CTNNB1	Wnt signalling, BMP induced genes[34]	rs30489
***WNT2B*, Wingless-type MMTV integration site family, member 2B**	Wnt ligand	Wnt signalling, BMP induced genes[34]	rs910697
WNT5A, Wingless-type MMTV integration site family, member 5A	Wnt ligand	Wnt signalling, BMP induced genes[34]	NA
WNT5B, Wingless-type MMTV integration site family, member 5B	Wnt ligand	Wnt signalling, BMP induced genes[34]	NA

Genes finally screened are depicted in bold.
NA denotes not available SNPs for a given gene considering our selection criteria. rs4444235 and rs9929218 are not shown, for they were included because of their previous associations and not because they fulfilled our functional criteria.
doi:10.1371/journal.pone.0012673.t003

Technologie AG, Baesweiler, Germany) in accordance with the manufacturer's instructions, at the Galician Public Fundation of Genomic Medicine in Santiago de Compostela. Cases and controls were extracted in mixed batches to avoid any kind of bias.

Candidate-gene selection

Both Wnt and BMP pathways were initially selected after the findings of Nishanian et al. [34], who demonstrated the interaction between these two pathways. Both pathways were thoroughly investigated through the Cancer Genome Anatomy Project site [35], but we failed to find any information regarding the BMP pathway in either this or other web browsers. For that reason, Wnt genes were selected by browsing the pathway through Biocarta [36], whereas BMP genes had to be strictly selected from previous literature [17,34]. Forty-one genes were finally selected to be included in the analysis.

SNP selection and genotyping

SNP selection criteria only considered functional markers with minor allele frequencies above 0.05 and at least two independent validation criteria as established in dbSNP [37]. This included all exonic variants selected with Pupasuite [38] and gene-regulatory regions in *cis* (5′or 3′ UTR ends), as defined by the FESD web browser [39]. 5′UTR variants were only included when they complied to the abovementioned criteria and were presumed to be in the potential binding site of a known transctiptional binding factor. 3′ UTR variants were included because of their potential relationship with miRNA binding regions [40]. Because some of the selected genes had no SNPs of such these kinds in any of the three browsers at the time of SNP selection, they ultimately had to be dropped out of the study. Finally, 43 SNPs were chosen within 21 genes to be screened as potential direct modifiers of CRC susceptibility (Table 3).

rs4444235 and rs9929218 are two variants lying in the near-by and intronic regions of *BMP4* and *CDH1*, respectively, that have been recently reported to be associated with the disease [8]. Considering that the SNPs that we had chosen within these two genes were not good taggers for these two variants (r-squared values were 0.6 for the SNPs in *BMP4*, and 0.02 for those in *CHD1*) (Figure 1), we decided to include them in our study as well, although they did not fulfill our selection criteria, making the total number of interrogated SNPs rise to 45.

Genotyping was performed with the MassARRAY (Sequenom Inc., San Diego, USA) technology at the Santiago de Compostela node of the Spanish Genotyping Center. Calling of genotypes was done with Sequenom Typer v4.0 software using all the data from the study simultaneously.

Statistical analyses

Quality control was performed, first by excluding both SNPs and samples with genotype success rates below 95%, with the help of the Genotyping Data Filter (GDF) [41]. Genotypic distributions for all SNPs in controls were consistent with Hardy-Weinberg equilibrium as assessed using a X^2 test (1^{df}). All p-values obtained were ≥ 0.05, thereby excluding the possibility of genotyping artifacts (data not shown). Population stratification was assessed with Structure v2.2 [42]. Briefly, the posibility of different scenarios was tested assuming a different number of underlying populations (k ranging from 1 to 4), allowing for a large number of iterations (25 K in the burn-in period followed by 500 K repetitions). The mean log likelihood was estimated for the data for a given k (referred to as L(K)) in each run. We as well performed multiple runs for each value of k computing the overall mean L(K) and its standard deviation. All results seemed to be concordant with the original assumption of a single existing population. Moreover, additional procedures for better confounding variable visualization were undertaken by means of a Principal Component Analysis (PCA) using the EIGENSOFT tool *smartpca* [43], although number of markers was very low. No differences were found of population stratification between cases and controls for either STRUCTURE or the first 10 components of the PCA analysis (Figure S2). After quality control 1746 samples (854 cases and 892 controls) and 37 SNPs remained for further analyses.

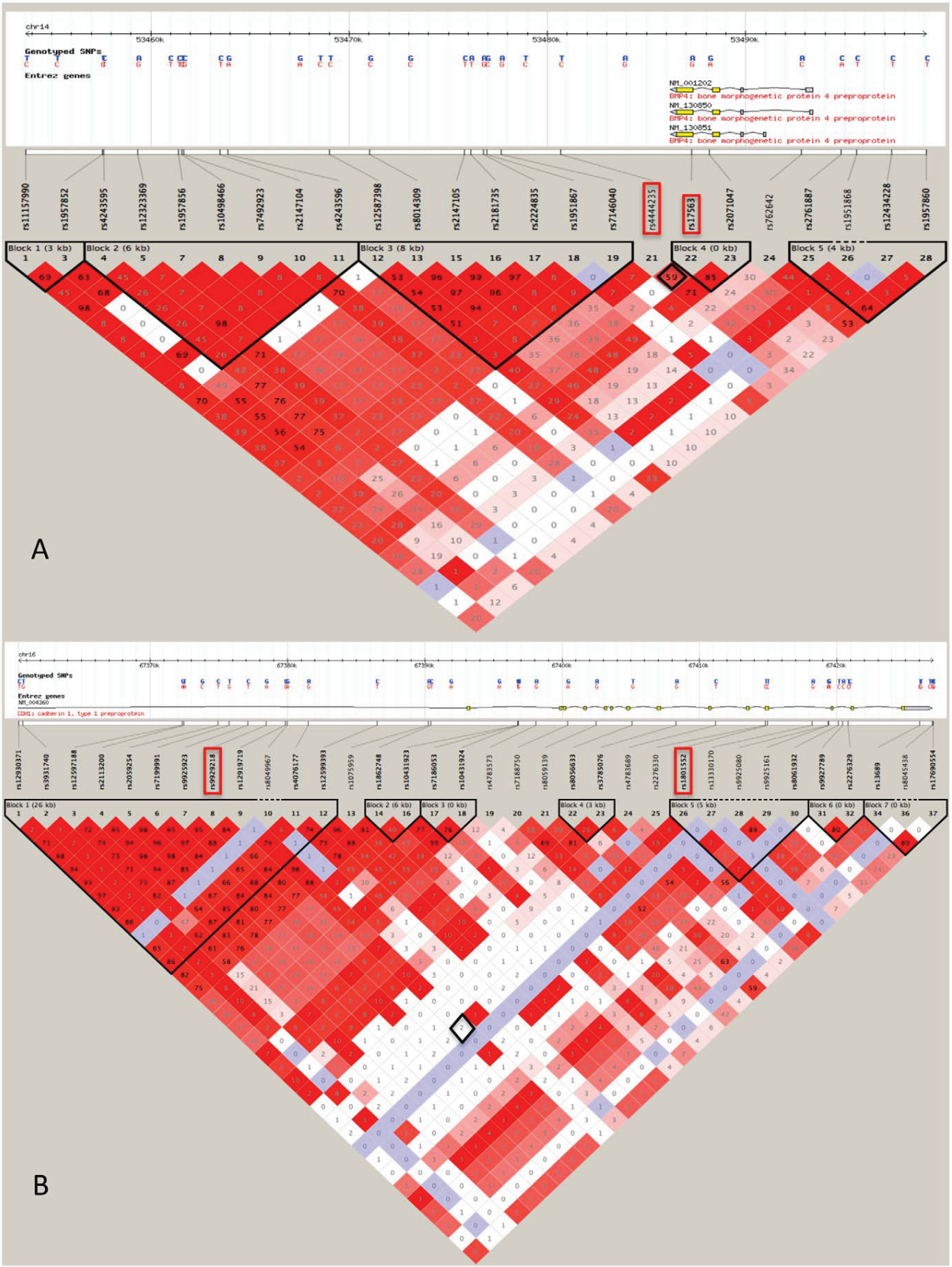

chr14
53460k
53470k
53480k
53490k
Genotyped SNPs
Entrez genes
NM_001202
BMP4: bone morphogenetic protein 4 preproprotein
NM_130850
BMP4: bone morphogenetic protein 4 preproprotein
NM_130851
BMP4: bone morphogenetic protein 4 preproprotein
rs11157990
rs1957852
rs4243595
rs12323369
rs1957856
rs10498466
rs7492923
rs2147104
rs4243596
rs12587398
rs8014309
rs2147105
rs2181735
rs2224835
rs1951867
rs7146040
rs4444235
rs17563
rs2071047
rs762642
rs2761887
rs1951868
rs12434228
rs1957860
Block 1 (3 kb)
Block 2 (6 kb)
Block 3 (8 kb)
Block 4 (0 kb)
Block 5 (4 kb)
A
chr16
67370k
67380k
67390k
67400k
67410k
67420k
Genotyped SNPs
Entrez genes
NM_004360
CDH1: cadherin 1, type 1 preproprotein
rs12930371
rs3931740
rs12597188
rs2113200
rs2059254
rs7199991
rs9925923
rs9929218
rs12919719
rs8049967
rs4076177
rs12599393
rs1075959
rs1862748
rs10431823
rs7186053
rs10431924
rs4783573
rs7188750
rs8059139
rs8056633
rs3785076
rs4783689
rs2276330
rs1801552
rs13330170
rs9925060
rs9925161
rs8061932
rs9927789
rs2276329
rs13689
rs8045438
rs17690554
Block 1 (26 kb)
Block 2 (6 kb)
Block 3 (0 kb)
Block 4 (3 kb)
Block 5 (5 kb)
Block 6 (0 kb)
Block 7 (0 kb)
B

Figure 1. Linkage disequilibrium blocks for the *BMP4* and *CDH1* genes. R-squared relationships between SNP pairs: A. rs4444235-rs17563 in *BMP4* and B. rs9929218-rs1801552 in *CDH1*.
doi:10.1371/journal.pone.0012673.g001

Association tests were performed by chi-squared tests for every single SNP and haplotypes where possible with both Haploview v4.0 [44] and Unphased [45]. In short, LD patterns across genes for which more than one SNP was genotyped were checked in Haploview and tested for association using Unphased (to check in any of the haplotypes was associated) and Haploview (to see which of the haplotypes was associated). Genotypic association tests, logistic regression analysis for sex and age adjustment, and stratified analysis between sporadic and familial groups were estimated with PLINK v1.03 [46]. OR and 95% confidence intervals were calculated for each statistic, and to address the issue of multiple-testing, permutation tests and the Bonferroni correction were used. Study power was estimated with CATS software [47].

Supporting Information

Figure S1 Haplotype structure and analysis for the 8 genes for which more than one SNP was genotyped. The table shows association values for each SNP generated by Haploview.
Found at: doi:10.1371/journal.pone.0012673.s001 (3.40 MB TIF)

Figure S2 Principal component analysis plot for the first vs. second component, comparing our case and control populations.
Found at: doi:10.1371/journal.pone.0012673.s002 (0.96 MB TIF)

Note S1
Found at: doi:10.1371/journal.pone.0012673.s003 (0.03 MB DOC)

Acknowledgments

We are sincerely grateful to all patients participating in this study who were recruited in 11 Spanish hospitals as part of the EPICOLON project. We thank Maria Magdalena-Castro (MMC), Olga Lortes (OL) and Eva Fernández (EF) for their excellent technical assistance.

Author Contributions

Conceived and designed the experiments: CRP. Performed the experiments: CFR. Analyzed the data: CFR ABF AA RMX SCB A. Carracedo. Contributed reagents/materials/analysis tools: LdC JC XB MA RJ XL A. Castells. Wrote the paper: CFR. Revised the manuscript and approved the final version: A. Castells SCB A. Carracedo.

References

1. Ferlay J, Bray F, Pisani P, Parkin DM (2004) GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide IARC Cancer Base No. 5. version 2.0.
2. Lichtenstein P, Holm NV, Verkasalo PK, Iliadou A, Kaprio J, et al. (2000) Environmental and heritable factors in the causation of cancer–analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 343: 78–85.
3. de la Chapelle A (2004) Genetic predisposition to colorectal cancer. Nat Rev Cancer 4: 769–80.
4. Castells A, Castellví-Bel S, Balaguer F (2009) Concepts in familial colorectal cancer: where do we stand and what is the future? Gastroenterology 137: 404–409.
5. Houlston RS, Peto J (2004) The search for low-penetrance cancer susceptibility alleles. Oncogene 23: 6471–6.
6. Chen SP, Tsai ST, Jao SW, Huang YL, Chao YC, et al. (2006) Single nucleotide polymorphisms of the APC gene and colorectal cancer risk: a case-control study in Taiwan. BMC Cancer 6: 83.
7. Naccarati A, Pardini B, Hemminki K, Vodicka P (2007) Sporadic colorectal cancer and individual susceptibility: a review of the association studies investigating the role of DNA repair genetic polymorphisms. Mutat Res 635: 118–45.
8. Houlston RS, Webb E, Broderick P, Pittman AM, Di Bernardo MC, et al. (2008) Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 40: 1426–35.
9. Tomlinson I, Webb E, Carvajal-Carmona L, Broderick P, Kemp Z, et al. (2007) A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 39: 984–8.
10. Zanke BW, Greenwood CM, Rangrej J, Kustra R, Tenesa A, et al. (2007) Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. Nat Genet 39: 989–94.
11. Broderick P, Carvajal-Carmona L, Pittman AM, Webb E, Howarth K, et al. (2007) A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 39: 1315–7.
12. Jaeger E, Webb E, Howarth K, Carvajal-Carmona L, Rowan A, et al. (2008) Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet 40: 26–8.
13. Tomlinson IP, Webb E, Carvajal-Carmona L, Broderick P, Howarth K, et al. (2008) A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 40(5): 623–30.
14. Tenesa A, Farrington SM, Prendergast JG, Porteous ME, Walker M, et al. (2008) Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 40(5): 631–7.
15. Cazier JB, Tomlinson I (2010) General lessons from large-scale studies to identify human cancer predisposition genes. J Pathol 220(2): 255–62.
16. Segditsas S, Tomlinson I (2006) Colorectal cancer and genetic alterations in the Wnt pathway. Oncogene 25: 7531–7.
17. Huang HC, Klein PS (2004) Interactions between BMP and Wnt signaling pathways in mammalian cancers. Cancer Biol Ther 3: 676–8.
18. Half E, Bercovich D, Rozen P (2009) Familial adenomatous polyposis. Orphanet J Rare Dis 4: 22.
19. Rowan AJ, Lamlum H, Ilyas M, Wheeler J, Straub J, et al. (2000) APC mutations in sporadic colorectal tumors: A mutational "hotspot" and interdependence of the "two hits". Proc Natl Acad Sci U S A 97: 3352–7.
20. Deng H, Ravikumar TS, Yang WL (2009) Overexpression of bone morphogenetic protein 4 enhances the invasiveness of Smad4-deficient human colorectal cancer cells. Cancer Lett 281: 220–31.
21. Deng H, Makizumi R, Ravikumar TS, Dong H, Yang W, et al. (2007) Bone morphogenetic protein-4 is overexpressed in colonic adenocarcinomas and promotes migration and invasion of HCT116 cells. Exp Cell Res 313: 1033–44.
22. Chen HM, Fang JY (2009) Genetics of the hamartomatous polyposis syndromes: a molecular review. Int J Colorectal Dis 24(8): 865–74.
23. Tao H, Shinmura K, Suzuki M, Kono S, Mibu R, et al. (2008) Association between genetic polymorphisms of the base excision repair gene MUTYH and increased colorectal cancer risk in a Japanese population. Cancer Sci 99: 355–60.
24. Schafmayer C, Buch S, Egberts JH, Franke A, Brosch M, et al. (2007) Genetic investigation of DNA-repair pathway genes PMS2, MLH1, MSH2, MSH6, MUTYH, OGG1 and MTH1 in sporadic colon cancer. Int J Cancer 121: 555–8.
25. Pomerantz MM, Ahmadiyeh N, Jia L, Herman P, Verzi MP, et al. (2009) The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. Nat Genet 41: 882–4.
26. Tuupanen S, Turunen M, Lehtonen R, Hallikas O, Vanharanta S, et al. (2009) The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. Nat Genet 41: 885–90.
27. Picelli S, Zajac P, Zhou XL, Edler D, Lenander C, et al. (2010) Common variants in human CRC genes as low-risk alleles. Eur J Cancer 46(6): 1041–8.
28. Slattery ML, Samowitz W, Ballard L, Schaffer D, Leppert M, et al. (2001) A molecular variant of the APC gene at codon 1822: its association with diet, lifestyle, and risk of colon cancer. Cancer Res 61: 1000–4.
29. Liu B, Zhang Y, Jin M, Ni Q, Liang X, et al. (2010) Association of selected polymorphisms of CCND1, p21, and caspase8 with colorectal cancer risk. Mol Carcinog 49: 75–84.
30. Hazra A, Fuchs CS, Chan AT, Giovannucci EL, Hunter DJ (2008) Association of the TCF7L2 polymorphism with colorectal cancer and adenoma risk. Cancer Causes Control 19: 975–80.
31. Ruiz-Ponte C, Vega A, Conde R, Barros F, Carracedo A (2001) The Asp1822Val variant of the APC gene is a common polymorphism without clinical implications. J Med Genet 38: E33.
32. Kemp Z, Thirlwell C, Sieber O, Silver A, Tomlinson I (2004) An update on the genetics of colorectal cancer. Hum Mol Genet 13 Spec No 2: R177–85.
33. Pinol V, Castells A, Andreu M, Castellvi-Bel S, Alenda C, et al. (2005) Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis colorectal cancer. JAMA 293: 1986–94.

34. Nishanian TG, Kim JS, Foxworth A, Waldman T (2004) Suppression of tumorigenesis and activation of Wnt signaling by bone morphogenetic protein 4 in human cancer cells. Cancer Biol Ther 3: 667–75.
35. The Cancer Genome Anatomy Project webpage (2006) http://cgap.nci.nih.gov/.
36. Biocarta (2006) http://www.biocarta.com/genes/index.asp.
37. 126.dbSNP (accessed 2007). http://www.ncbi.nlm.nih.gov/projects/SNP/.
38. Conde L, Vaquerizas JM, Dopazo H, Arbiza L, Reumers J, et al. (2006) PupaSuite: finding functional single nucleotide polymorphisms for large-scale genotyping purposes. Nucleic Acids Res 34: W621–5.
39. Kang HJ, Choi KO, Kim BD, Kim S, Kim YJ (2005) FESD: a Functional Element SNPs Database in human. Nucleic Acids Res 33: D518–22.
40. Jackson RJ, Standart N (2007) How do microRNAs regulate gene expression? Science (367): re1.
41. Genotyping Data Filter Software (2009) http://bioinformatics.cesga.es/gdf/nav_input.php.
42. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–59.
43. Patterson N, Price AL, Reich D (2006) Population structure and Eigenanalysis. PLoS Genet 2(12.
44. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–5.
45. Dudbridge F (2003) Pedigree disequilibrium tests for multilocus haplotypes. Genet Epidemiol 25: 115–21.
46. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, et al. (2007) PLINK: a toolset for whole-genome association and population-based linkage-analysis. American Journal of Human Genetics; 81: 559–75. http://pngu.mgh.harvard.edu/~purcell/plink/.
47. Skol AD, Scott LJ, Abecasis GR, Boehnke M (2006) Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 38: 209–13.

Chapter 3:
A Colorectal Cancer Genome-Wide Association Study in a Spanish cohort identifies a new colorectal cancer susceptibility variant at 8p12

Manuscript in preparation

A Colorectal Cancer Genome-Wide Association Study in a Spanish cohort identifies a new colorectal cancer susceptibility variant at 8p12

Fernandez-Rozadilla C[1], Cazier JB[2], Carvajal-Carmona L[2], Palles C[2], Abulí A[3], Bujanda L[4], Clofent [5], Moreno V[6], Lamas MJ[7], Baiget M[8], López JL[9], Andreu M[3], Bessa X[3], Jover R[10], Llor X[11], Castells A[12], Tomlinson I[2]. Castellví-Bel S[12], Carracedo A[1], Ruiz-Ponte C[1], for the EPICOLON consortium

[1]Galician Public Fundation of Genomic Medicine (FPGMX)-Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER)-Genomics Medicine Group-Hospital Clínico Santiago de Compostela-University of Santiago de Compostela. Spain.; [2]Wellcome Trust Centre for Human Genetics, University of Oxford, UK; [3]Gastroenterology Department, Hospital del Mar, Barcelona, Spain; [4]Colorectal Cancer Multidisciplinary Unit, Donostia Hospital, University of the Basque Country, San Sebastián, Spain; [5]G; [6]Service of Oncology and Cancer Registry, Catalan Institute of Oncology (ICO); [7]Oncology Pharmacy Unit, Complejo Hospitalario Universitario of Santiago (CHUS), Spain; [8]Molecular Genetics Unit, Hospital de Santa Creu I Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain; [9]Pharmacogenetics & Pharmacogenomics Laboratory, Pharmacy Unit, Hospital General Universitario Gregorio Marañón, Madrid, Spain; [10]Gastroenterology Department, Hospital General de Alicante, Alicante, Spain; [11]Section of Digestive Diseases and Nutrition, University of Illinois at Chicago, Chicago, IL, USA;[12]Department of Gastroenterology, Hospital Clínic, CIBERehd, IDIBAPS, University of Barcelona, Barcelona, Spain; for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association

Colorectal cancer is a known to be a complex disease, with much of the expected inherited risk being due to several common low risk variants. Genome-Wide Association Studies (GWAS) have conveniently identified 14 loci harbouring some of the susceptibility variants that influence the risk of developing CRC. Nevertheless, these have only been able to explain part of the missing heritability, with the remaining yet to be discovered. We followed a GWAS approach and performed a genome-wide association study in a Spanish cohort of 881 cases and 667 controls. After association analyses, 64 variants on 24 genomic loci were found to be associated with CRC with p-values in the order of 10^{-5}. No evidences of association in the nearby regions of any of these variants in the CORGI British cohort were found. However, there were evidences for 8 of these loci that minor allele frequencies (MAFs) between Northern and Southern-European populations may be different. Based on this, we evaluated the association signals of these eight loci in a Spanish replication cohort of 1481 cases and 1850 controls. One of these SNPs, rs11987193 at 8p12 was positively replicated (pooled p=2.061x10^{-5}). The T allele of this SNP shows a protective effect on CRC risk and may be related to *DUSP4* function.

Keywords: colorectal cancer, GWAS, SNPs, Spanish cohort, risk variants

Introduction

Even though genetic susceptibility is thought to be responsible for almost 35% of al CRC cases[1], high penetrance mutations in Mendelian predisposition genes, such as *APC*, the mismatch repair (MMR) genes, or *MUTYH* have only been able to explain <5% of CRC cases[2]. The recent advances in the field of genetic epidemiology have validated the hypothesis that at least part of that remaining missing susceptibility lies in the form of multiple common low-risk variants, each conferring a modest effect on disease risk.

GWAS are one of the most widespread methodologies for the detection of such susceptibility loci. The procedure (in opposition to gene-candidate association studies) offers an unbiased strategy for the detection of new low-penetrance variants, for it does not assume any *a priori* hypothesis on the location of these loci. This advantage has been proved important, since so far this kind of surveys have successfully identified variants at 8q24.21 (rs6983267)[3], 8q23.3 (rs16892766), 10p14

(rs10795668)[4], 11q23 (rs3802842)[5], 15q13 (rs4779584)[6], 18q21.1 (rs4939827)[7], 14q22.2 (rs4444235), 16q22.1 (rs99292218), 19q13.1 (rs10411210), 20q13.3 (rs961253)[8], 1q41 (rs6691170 and rs6687758), 3q26.2 (rs10936599), 12q13.1 (rs11169552 and rs7136702) and 20q13.33 (rs4925386)[9]. The combined effect of the variants at these 14 loci altogether is thought to explain ~7% of the familial cancer risk[9]. Still, there is a high proportion of the CRC cases for which no genetic cause has been identified.

In this study we have attempted a new screening for CRC susceptibility variants by undertaking a GWAS approach on our cohort of 881 CRC cases and 667 controls from the Spanish population. The use of a Southern-European dataset is also a novelty; since all of the populations were GWAS analyses have been conducted so far have been of Northern-European origin. This would provide additional confirmation of the relationship of the 14 described loci. We must however also consider the possibility that there may be differences, at these or other particular loci in the genome, between these sets of populations, which could effectively lead to discrepancies in the tagging of the real causative variants behind the association signals.

Materials and Methods

Study populations. Subjects on Phase I were 881 cases and 473 controls ascertained through the EPICOLON II Project and 194 additional controls from the Spanish National DNA bank. The EPICOLON Consortium comprises a prospective, multicentre and population-based epidemiology survey of the incidence and features of CRC in the Spanish population[10,11]. Cases were selected as patients with *de novo* histologically confirmed diagnosis of colorectal adenocarcinoma. Patients with Familial Adenomatous Polyposis, Lynch Syndrome or Inflammatory Bowel Disease-related CRC, and cases where patients or family refused to participate in the study were excluded.
Median age for cases on stage I was 73 (range 26-95), whereas mean was 71.2 years. Hospital-based controls were recruited together with cases for the EPICOLON initiative. All of these were confirmed to have no cancer or prior history of neoplasm and no family history of CRC. Controls were randomly selected and matched with cases for hospital, sex and age (± 5 years). Controls from the National DNA bank were also genotyped, to lessen the deficit in controls. They were matched for sex, age (± 10 years) and geographical origin of the sample with the remaining cases. Both cases and controls were of European ancestry and from Spain (stated, where possible, as all four grandparents being Spanish). Gender and hospital distribution of samples for case and control groups is shown on Supplementary Table 1.
Samples on Phase II consisted of: 1436 CRC patients and 1780 controls. Of these, 821 CRC patients were recruited in 4 different Spanish centers: Hospital Sant Pau, Hospital Gregorio Marañón, Catalan Institute of Onconlogy (ICO) and the CHUS hospital in Santiago de Compostela, 105 CRC cases and 1330 controls came from the Spanish National DNA bank, and 510 CRC cases and 450 DNA controls belonged to the EPICOLON I Project. Of these, 60.4% were male and 39.6% female; age median was 69.61 (69.02-70.20) for cases and 52.00 (51.42-52.58) for controls.
DNA was obtained from frozen peripheral blood by standard extraction procedures for all samples. Cases and controls were extracted in mixed batches to avoid bias.

Ethics statement. The study was approved by the "Comité Ético de Investigación Clínica de Galicia", and each of the institutional review boards of the hospitals where samples were collected. All samples were obtained with written informed consent reviewed by the ethical board of the corresponding hospital, in accordance with the tenets of the Declaration of Helsinki.

SNP genotyping and QC. Affymetrix chip 6.0 (Ayymetrix. CA USA) was chosen to obtain genome-wide coverage for our SNP susceptibility scan in phase I genotyping. The chip includes probes for almost 1M SNP markers. Genotyping in phase 2 was conducted by Sequenom MassARRAY technology (Sequenom Inc.. San Diego. CA. USA). Genotyping for both stages was performed at the Santiago de Compostela node of the Spanish Genotyping Center. Genotype calling for the Affymetrix 6.0 array intensities was performed with the Birdseed algorithm, included within Birdsuite v1.4[12]. Samples were organised in 23 batches of $16<n<99$ according to hospital of origin for computational purposes. We obtained valid genotypes for 909.622 SNPs by these means. Conversion of genotype data into PLINK v1.07 format[13] was performed using in-

house scripts. Quality control of the data included the removal of both loci and samples with genotyping success rates <99% and concordance check of genders between clinical recorded data and Affymetrix asigned sex. Hardy-Weinberg equilibrium (HWE) was evaluated with a 1 degree of freedom (df) X^2 test, or Fisher´s exact when genotype counts<5; markers $<1x10^{-4}$ threshold on the unaffected group of samples were removed from further analyses. SNPs with MAFs below 0.05 were also eliminated due to power-related reasons and to avoid unnecessary noise signals. Differential missingness between cases and controls was also accounted for by excluding markers with p-values below $1x10^{-4}$. A total of 674.718 SNPs remained after this filtering.

To address the possibility of underlying population stratification, Principal Component Analysis (PCA) on a set of 98,986 independent SNPs (maximum pairwise r-squared value of 0.1) was also performed on the cohort with the help of the EIGENSOFT *smartpca* tool[14]. Long-range LD regions, as described by Price *et al*.[15], were also removed from this analysis. Results for this PCA are depicted on Figure 1A. Outliers (taken as samples with >5% distance from the cluster centroid), as well as samples spread on principal components 1 and 2 (Eigenvector 1 <-0.01 and Eigenvector 2 >0.05) were removed from subsequent analyses, since they deviated from the main cloud. No evidences were found of population stratification between cases and controls for the first 10 components of the PCA analysis (Figure 1B). Other potentially confounding variables, such as Nsp-vs-Sty-genotyped markers, hospital of collection, genotyping plate, or geographical origin of the samples were also checked for as sources for stratification (data not shown). All results seemed to be concordant with the original assumption of a single existing population except for hospital of origin. When considered as a confounding variable, the EPICOLON cohort clustered into three separate subgroups: samples from the Donostia hospital (VAS dataset), the only collection centre for the Basque Country regions (North of Spain), samples from the Meixoeiro hospital (GAL dataset), the single collection point in Galicia (NW Spain), and all others (REST dataset) (Figure 1C). An additional PCA with the EPICOLON II cohort and the HapMap3 populations was also performed to illustrate the clustering of the EPICOLON II cohort with the HapMap3 populations[16] (Figure 1D). The plotting was then restricted to those with Caucasian origin (CEU and TSI) or those with a potential Spanish contribution (MEX) (Figure 1E). Samples that clustered away from the European end of the plot (showing evidences of non-European ancestry) were also excluded from further analyses.

The final dataset was comprised of 1477 samples (848 cases and 629 controls). The total count per subgroup was 167 for VAS, 366 for GAL and 944 for REST.

A second stage of quality control was undertaken after the association analyses by means of the Evoker software[17]. Associated SNPs at a selected threshold were plotted to compare the efficiency of the calling procedure by comparing the intensity clusters derived from the genotyping array against the genotype clusters assigned by the calling algorithm.

Statistical analysis. Association analysis was assessed as a 1 df X^2 allelic test for each of the three subgroups independently for phase one, and for second stage replication, with the help of PLINK[13]. The adequacy of the distribution of p-values was evaluated using quantile-quantile (Q-Q) plots of test statistics and lambda genomic inflation factors. Meta-analysis was conducted using the R package META[18] and PLINK. Both methods are based on a Mantel-Haenszel approach for data pooling. Cochran´s Q statistic[19] and the I^2 heterogeneity index[20] were also estimated to account for inter-population heterogeneity. Risks associated with each marker were estimated by odds ratios (OR) and their 95% confidence intervals (CIs), assuming both fixed and random-effect models.

Imputation of the linkage disequilibrium (LD) blocks around each of the 24 loci that showed evidences of association was accomplished with Impute v2[21] using two reference panels: 1000 Genomes Project (b36) for wide coverage[22], and HapMap3 (r2 b36) for deep coverage[16]. Imputation results were filtered by minor allele frequency (MAF) of the markers, since the procedure generates genotypes for a high number of rare variants that could give spurious association results (thus SNPs with MAFs<5% were excluded), missing data proportion (set to a 5% max), and the *frequentist-add-proper-info* column of the output. This latter proportion is indeed the ratio of the empirically observed variance of the allele dosage to the expected binomial variance p(1-p) at HWE, where p is the observed allele frequency from HapMap[23]. Optimal values should be within the (0.4-1) range and provide a measure for quality and accuracy of the imputation. Since the proportion of cases and controls deviates significantly from the standard 1:1, we also considered the possibility that the genotype probabilities for each marker were different in both subsets.

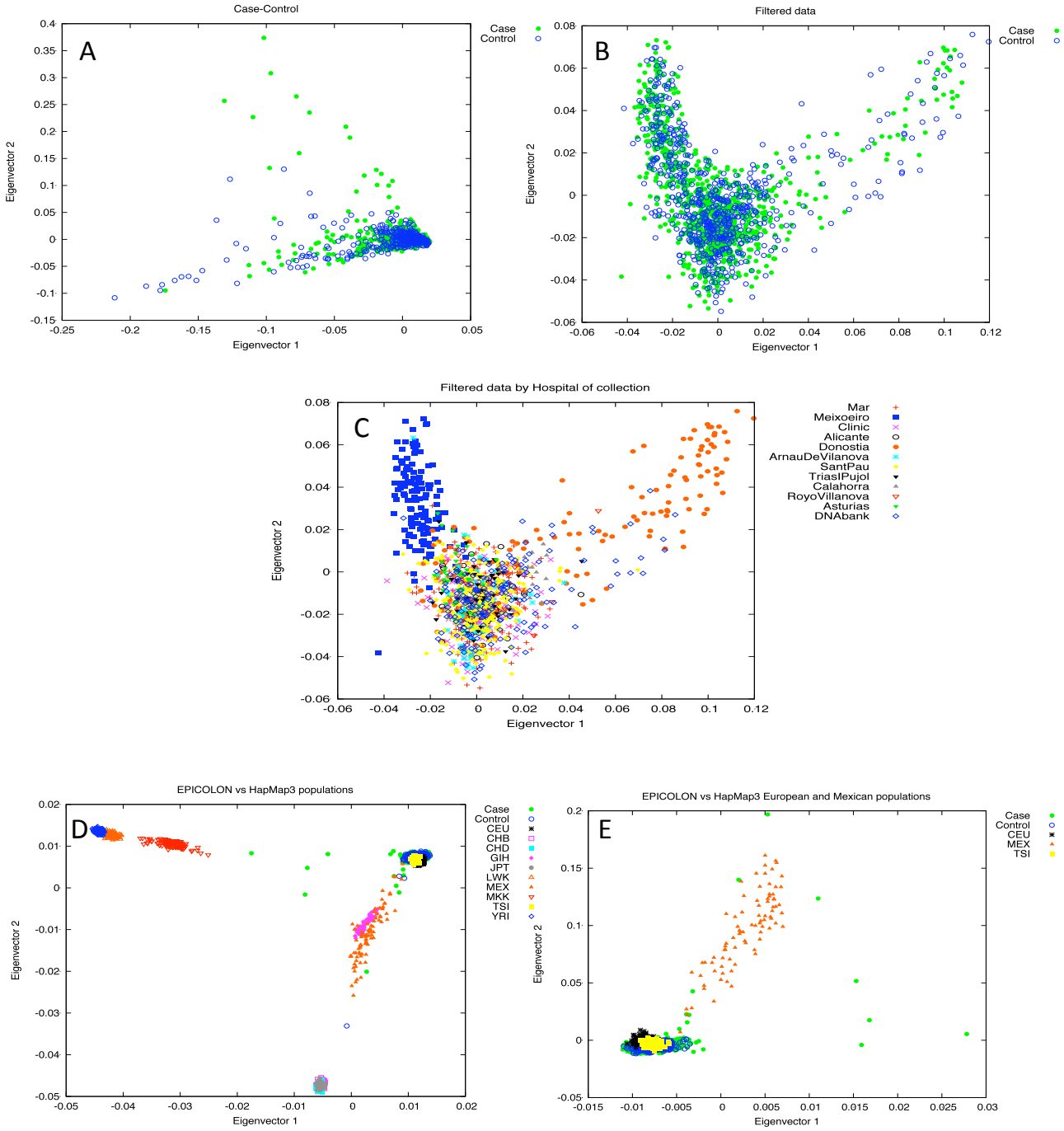


Figure 1. PCA analysis on the EPICOLON cohort. A. Raw data; B: filtered data by case/control; C: filtered data by hospital of origin; D: EPICOLON and HapMap3 populations; E: EPICOLON. CEU. TSI and MEX. Significant differences may be seen in section C. with Meixoeiro and Donosti hospitals deviating from the main cloud.

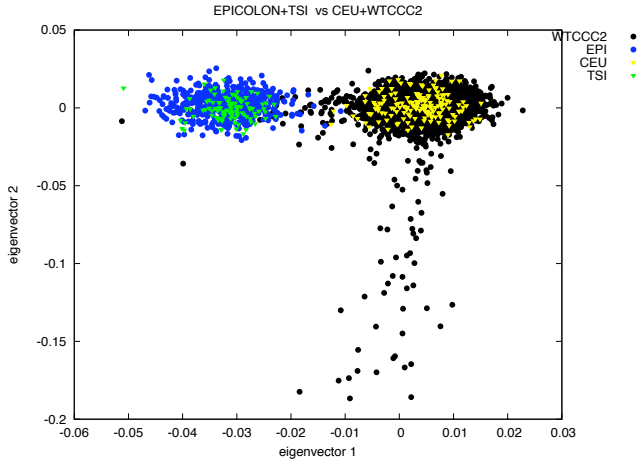


Figure 2: PCA analysis on the WTCCC (Affymetrix 6.0 data), HapMap3 CEU and TSI and EPICOLON populations. A set of 15,000 independent markers was used to perform the analysis. The first eigenvector seems to separate the Northern and Southern European populations.

Thus we filtered out SNPs for which the probability of two out of the three genotypes was ≥25% in at least 5% of the cases or the controls. Pooled analysis was performed by logistic regression with stage and subgroup as covariates. Additional statistical calculations, such as Pearson´s product-moment correlation values, were calculated using R. Imputation results were plotted with the help of SNAP[24].

Results

We observed during our quality control check-up procedure, that there was an important batch effect due to differences by hospital of collection of the sample, dividing the EPICOLON cohort into three separate subgroups. Pondering this, we considered it appropriate to proceed on forward with the association analyses by contemplating each cluster as a separate population (the GAL, VAS and REST groups), in order to avoid any bias leading to an increased false positive association rate.

Association results were thus obtained for each of the GAL, VAS and REST subpopulations separately. Q-Q plots for the subgroups showed some signs of inflation in the distribution of the association p-values for the GAL and REST groups (Supplementary Figures 1A and 1B. respectively). The VAS group however, showed better fitting but no deviations indicative of association hits, probably due to its smaller sample size (Supplementary Figure 1C). Lambda genomic factor calculations (1.04192, 1.02323 and 1.02292 for the GAL, VAS and REST populations, respectively) were however consistent with no evidences of an increased false discovery rate.

Meta-analysis from the association results in the three separate populations was carried out with the R META package as well as PLINK. Q-Q plot distribution of these results is depicted on Supplementary Figure 1D. Ninety-seven percent of the SNPs showed consistency in p-value calculations (differences <0.1 between both approaches); 0.43% showed differences between 0.1-0.15, and 2.11% of the markers showed a discrepancy of 0.15 or greater in the p-values obtained by either method (Supplementary Figure 2). Sensitivies for both methods were similar with only a 13% increased loci detection for META at a p-value threshold of $1x10^{-4}$. However, Heterogeneity between the GAL, VAS and REST groups was defined as I^2>75%. For markers above this threshold, a random-effects model was considered, whereas fixed-effect results were reported otherwise.

With these criteria, we found 93 SNPs associated at a level of $1x10^{-4}$ or below. Evoker intensity plots were created to allow for a visualisation of the intensity clusters in comparison with the genotypes assigned by the calling algorithm in order to detect potential calling artefacts (Supplementary Figure 3). Sixty-four final SNPs located on 24 different genomic loci remained after this checking.

The locations of these 64 SNPs were fine-mapped by imputing the LD blocks around these regions, in order to refine the association signals. Table 1 provides with a summary of the loci and extent of the imputed regions. We approached this analysis in two different ways: a) imputation of the LD blocks in all samples altogether, then splitting of the dataset into the three subgroups, performance of the association analyses and meta-analysis of these; b) imputation of each chromosomal region separately in every subgroup, running of the association tests and meta-analysis. Results for both were however similar, as ascertained by correlation analysis of the p-values at every segment (correlation coefficients 0.876-0.998). An example of the correlation plots for one of the regions may found on Supplementary Figure 4. This analysis improved the association at loci 1p33 (best SNP rs12060081), 14q31.3 (rs2057115), 15q21.3 (rs7176932) and 22q12.3 (rs17725348) (Supplementary Figure 5).
Evidences of association (taken as the presence of markers with p-values below $1x10^{-4}$) were screened for in the ±1Mb region surrounding every SNP in the CORGI GWAS[3]. Only 5 of these locations showed to have some CORGI associated SNP at the established threshold, but r-squared pairwise measures of LD evidenced them all to be independent signals ($r^2<0.8$ for all-data not shown).

MAFs in controls for all 64 associated SNPs were then checked in the HapMap3 CEU and TSI populations, as well as the Wellcome Trust Case Control Consortium (WTCCC2) control cohorts[25]. Ten SNPs in 8 loci showed significant and consistent deviations in TSI and EPICOLON MAFs compared to the Northern-European

Table 1: Associated loci and imputation regions. Location of the 24 associated loci and description of the regions that were imputed for finer mapping.

CHR	LOCUS	IMPUTATION REGION
1	1p33	47,985,000-48,255,000
2	2p25.2	5,518,000-5,623,000
2	2p24.1	22,284,000-22,608,000
3	3p21.31	46,887,500-47,562,000
3	3q12-q13	120,991,750-121,488,000
5	5q35.1	172,706,000-172,719,000
6	6q16.1	99,253,000-99,325,000
6	6q23.1-q23.2	131,177,000-131,499,500
8	8p12	29,383,500-29,403,500
8	8q13.3	72,696,800-72,704,400
8	8q22.1	96,663,000-96,777,000
10	10p15.1	5,619,000-5,736,000
10	10q23.31	92,677,000-92,789,200
12	12q24.31	119,533,000-119,597,000
13	13q32.3	99,584,500-99,846,500
14	14q31.3	85,090,000-85,120,800
14	14q32.12	92,202,500-92,271,850
15	15q21.3	52,143,300-52,192,400
15	15q25.3	86,171,000-86,253,700
17	17p13.2	13,145,300-13,256,000
17	17p12	5,201,500-5,322,700
18	18p11.22	8,552,000-8,595,800
18	18q21.2	50,980,500-51,327,000
22	22q12.3	34,479,200-34,790,200

populations (CEU and WTCCC2) (Table 2). PCA analysis on the four populations and 15,000 independent SNPs effectively separated the Northern and Southern-European populations (Figure 2). Given this evidence, we decided to replicate the best-associated markers at these loci (taken as either directly genotyped or imputed best score SNP) in an independent
Spanish cohort. rs7087402 at 10q23.31 could not be included for genotyping design reasons.

One of these SNPs, rs11987193 was successfully replicated in the second stage (p=0.039, OR=0.847 (0.725-0.991); Table 3). Although the association signal was modest, pooled analysis of the data from both stages was consistent with the presence of a CRC susceptibility variant in this location ($p=2.061x10^{-5}$; OR=0.788

Table 2. MAF comparison for the associated SNPs. Frequencies for all 64 SNPs at 24 associated loci for the EPICOLON, WTCCC2 control cohorts, CEU and TSI HapMap3 populations. Loci with consistent deviations in frequencies between Northern and Southern European populations are highlighted in bold.

LOCUS	SNP	POSITION	MAF CEU (N=180)	MAF WTCCC2 (N=5380)	MAF EPICOLON* (N=625)	MAF TSI (N=102)
1p33	**rs12080929**	**48,208,735**	**0.265**	**0.248**	**0.312**	**0.324**
2p25.2	rs4669394	5,541,078	0.084	0.073	0.08	0.103
2p24.1	rs1554266	22,284,300	0.425	0.454	0.373	0.426
2p24.1	rs1554267	22,284,451	0.42	0.454	0.371	0.415
2p24.1	rs1554269	22,284,627	0.42	0.454	0.372	0.421
2p24.1	rs4557006	22,297,345	0.425	0.455	0.373	0.436
2p24.1	rs6759922	22,303,754	0.42	0.450	0.370	0.436
2p24.1	rs4416248	22,309,026	0.42	0.451	0.370	0.436
3p21.31	rs10461018	46,970,246	0.469	0.419	0.454	0.426
3p21.31	rs2061197	46,976,354	0.465	0.419	0.454	0.426
3p21.31	rs749511	47,010,739	0.447	0.407	0.436	0.422
3p21.31	rs2305634	47,018,542	0.46	0.422	0.472	0.429
3p21.31	rs2278963	47,029,873	0.447	0.406	0.436	0.422
3p21.31	rs7610636	47,039,440	0.465	0.425	0.468	0.441
3p21.31	rs11917361	47,045,501	0.447	0.405	0.435	0.422
3p21.31	rs6442055	47,085,726	0.447	0.406	0.436	0.422
3p21.31	rs6767907	47,137,665	0.447	0.406	0.436	0.422
3p21.31	rs9837343	47,152,392	0.447	0.406	0.435	0.422
3p21.31	rs295442	47,310,885	0.434	0.403	0.457	0.436
3p21.31	rs17410853	47,338,975	0.398	0.364	0.401	0.377
3p21.31	rs8180040	47,363,951	0.434	0.402	0.454	0.431
3p21.31	rs4858888	47,380,309	0.434	0.403	0.454	0.431
3p21.31	rs2062278	47,391,765	0.429	0.396	0.455	0.431
3p21.31	rs12636851	47,438,571	0.429	0.395	0.446	0.426
3p21.31	rs6800271	47,445,791	0.429	0.3970	0.451	0.431
3p21.31	rs3816779	47,518,393	0.429	0.399	0.453	0.436
3p21.31	rs7628631	47,535,867	0.429	0.396	0.445	0.436
3p21.31	rs11130137	47,553,842	0.429	0.398	0.449	0.436
3q12-q13	rs2472680	121,010,466	0.031	0.049	0.072	0.034
3q12-q13	rs6438550	121,019,507	0.031	0.050	0.072	0.025
5q35.1	rs11745626	172,706,309	0.181	0.156	0.186	0.113
5q35.1	rs11740081	172,707,280	0.183	0.157	0.188	0.113
5q35.1	rs17733311	172,712,710	0.181	0.163	0.193	0.108
6q16.1	rs12213685	99,288,865	0.146	NA	0.107	NA
6q16.1	rs4262197	99,299,694	0.128	0.149	0.105	0.127
6q16.1	rs6941632	99,302,147	0.155	0.172	0.119	0.152
6q16.1	rs6936798	99,305,520	0.155	0.172	0.119	0.152
6q16.1	rs7750336	99,322,459	0.155	0.172	0.121	0.152
6q16.1	rs9398904	99,323,739	0.119	0.145	0.103	0.118
6q16.1	rs7740725	99,324,762	0.119	0.144	0.104	0.118
6q23.1-q23.2	rs12199765	131,192,418	0.288	0.257	0.202	0.275
8p12	**rs11996339**	**29,386,099**	**0.385**	**0.385**	**0.448**	**0.515**
8p12	**rs11987193**	**29,391,927**	**0.283**	**0.267**	**0.309**	**0.309**
8p12	**rs12548021**	**29,400,381**	**0.35**	**0.392**	**0.289**	**0.328**
8q13.3	rs17788534	72,697,475	0.115	0.139	0.141	NA
8q22.1	rs3104964	96,664,912	0.415	0.400	0.367	0.497

** Only control samples from our EPICOLON population were considered for MAF calculations.*

Table 2. (Continuation).

LOCUS	SNP	POSITION	MAF CEU (N=180)	MAF WTCCC2 (N=5380)	MAF EPICOLON (N=625)	MAF TSI (N=102)
10p15.1	**rs7074607**	**5,623,371**	**0.155**	**0.147**	**0.102**	**0.132**
10q23.31	**rs7087402**	**92,760,125**	**0.492**	**0.495**	**0.444**	**NA**
12q24.31	rs568489	119,578,624	0.434	0.393	0.397	0.48
12q24.31	rs2686555	119,579,555	0.434	0.392	0.404	0.48
13q32.3	rs17196583	99,624,356	0.17	0.176	0.145	0.206
14q31.3	rs7148493	85,094,169	0.378	0.380	0.363	0.391
14q32.12	**rs8177528**	**92,247,404**	**0.364**	**0.362**	**0.335**	**0.343**
15q21.3	rs1897019	52,163,314	0.362	0.397	0.313	0.422
15q21.3	rs4644815	52,163,793	0.362	0.397	0.313	0.422
15q21.3	rs4644804	52,164,106	0.369	0.398	0.316	0.422
15q21.3	rs12913167	63,265,330	0.365	0.397	0.315	0.426
15q25.3	rs16941001	86,249,170	0.067	0.081	0.059	0.074
17p13.2	rs12603094	5,288,680	0.137	0.139	0.103	0.172
17p13.2	rs16954697	5,297,637	0.124	0.132	0.098	0.191
17p12	rs9898623	13,255,126	0.08	0.076	0.085	0.074
18p11.22	**rs10502376**	**8,579,765**	**0.388**	**0.478**	**0.496**	**NA**
18q21.2	**rs2958182**	**51,300,019**	**0.319**	**0.337**	**0.261**	**0.284**
22q12.3	**rs956119**	**34,582,213**	**0.062**	**0.073**	**0.104**	**0.083**

** Only control samples from our EPICOLON population were considered for MAF calculations.*

Table 3. Association results for stage II. P-values and ORs for the replication stages.

LOCUS	SNP	P-VALUE PHASE I	P-VALUE PHASE II	OR PHASE II
1p33	**rs12080061***	1.620E-05	0.083	0.869 (0.7414-1.019)
8p12	**rs11996339**	9.697E-08	0.690	0.971 (0.842-1.12)
8p12	**rs11987193**	9.750E-06	0.039	0.847 (0.724-0.992)
8p12	**rs12548021**	1.071E-06	0.234	1.095 (0.9431-1.271)
10p15.1	**rs7074607**	8.751E-05	0.174	0.867 (0.705-1.065)
10q23.31	**rs7087402**	5.180E-06	NA	NA
14q32.12	**rs8177528**	5.471E-05	0.423	1.056 (0.910-1.225)
18p11.22	**rs10502376**	9.819E-05	0.986	0.973 (0.842-1.125)
18q21.2	**rs2958182**	5,657E-05	0.714	1.001 (0.861-1.164)
22q12.3	**rs17725348***	4.011E-05	0.273	0.875 (0.690-1.110)

** Denotes imputed SNPs; NA: not available*

Table 4. Replication results for the already-described loci. Half of the loci showed direct evidences of association

LOCUS	REPORTED SNP	ASSOCIATED SNP IN EPICOLON	R^2	P-VALUE	OR (95% CI)
8q24	rs6983267	rs6983267	-	0.065	0.871 (0.751-1.009)
10p14	rs10795668	rs10905436	0.929	0.066	0.623 (0.737-1.010)
11q23	rs3802842	rs3802840	1	0.037	1.190 (1.010-1.402)
12q13	rs11169552	rs11169567	0.166	9.502E-04	1.282 (1.106-1.486)
15q13	rs4779584	rs4779584	-	8.772E-04	1.389-1.144-1.686)
18q21	rs4939827	rs7226855	0.95	8.204E-03	0.820 (0.707-0.950)

(0.706-0.879)). The other two SNPs at this locus, rs11996339 and rs12548021 did not appear as significant in this second stage, although this was pretty much expected, since the signals were independent.

Aside from the search of new susceptibility variants, we also investigated the association signals for the 14 known CRC susceptibility loci. Direct evidence of replication (taken as the presence of an associated SNP with p-value <0.1 near the described location) was found for 6 of the sites (see Table 4). Imputation of the LD regions around these associated loci was conducted to search for an enhancing of the signals. No significant improvements were found, except for locus 15q13 (data not shown). Association results and MAF measures for all 16 SNPs at the 14 loci (considered as the result obtained for the best matching proxy) are visualised on Table 5 and compared to the described literature.

Discussion

Genome-wide association studies have so far successfully identified 14 susceptibility-to-colorectal-cancer loci[8,9]. Although this has been a significant improvement in the unravelling of the genetic basis of the disease, these variants alone do not completely explain all the inherited variation that has been attributed to CRC.

Following the lead of the previous studies, we addressed the issue of trying to detect new colorectal cancer susceptibility variants through the performance of a GWAS in a Spanish cohort. This was the first attempt to perform a CRC GWAS in a Southern European population. By these means, we were able to positively identify a new susceptibility variant, rs11987193, at 8p12.

During the analysis, we were faced with the fact that, although there were no differences in case and control populations, there was a significant stratification issue determined by the hospital of origin of the samples. Because of this, the analyses had to be modified to match our case scenario without losing significant power. The subdivision of the population had also great implications on the imputation procedure, although in this case both of the approaches taken seemed to give concordant results (as seen from the correlation analysis of the p values). This is most likely a result from the three subgroups having different sizes, with the largest being the determinant in the outcome of the analysis.

Nevertheless, the substructure in our cohort did not seem to greatly affect outcome quality. The evaluation on the already-described signals achieved direct replication for 6 of the loci (8q24, 10p14, 11q23, 12q13, 15q13 and 18q21), although the best-associated markers for these regions did not always match with the best proxy for the already described SNPs. This would make sense if we consider that any given GWAS relies on an indirect approach, and we would expect the associated SNPs to only be taggers of the real causative variant. Results for allele frequencies and ORs seem consistent with the bibliography[8,9].

In a similar way, we carried out additional quality control procedures during our study. We tested the performance of two different software at the meta-analysis step: the R package META and PLINK´s own implementation. The former showed higher sensitivity rates but the computational intensity of the method does not favour its use in meta-analysis studies. Evoker plots were also examined, and a considerable proportion of the association signals (31%) was by these means identified as artefacts generated during the batch calling procedure. This was particularly true for SNPs with lower allele frequencies, for the proportion of homozygous individuals for the minor allele would be low enough in batches with lower sample sizes for these individuals to be wrongly called as heterozygous. Hence, we encourage for this additional controls to be performed in order to reduce false positive findings.

The association analysis in itself provided with positive results in 64 SNPs on 24 different genomic loci at a p-value <0.0001. A first attempt at replication was aimed by inspection of these association signals on the CORGI cohort[3]. However, none of the signals seemed to be shared between datasets. This lack of replication could be

Table 5. Data for the 14 reported loci. Comparison between bibliography data and EPICOLON association results for the 16 SNPs at the 14 susceptibility loci.

SNP[REF]	LOCUS	REP ALLELE	REP MAF CTRLS	REP ALLELIC OR (95% CI)	BEST PROXY AFFY 6.0	R^2	MINOR ALLELE	MAF	OR (95% CI)	EPICOLON P-VALUE
rs6687758[9]	1q41	G	0.2	1.09 (1.06-1.12)	rs6691195	1	T	0.19	1.104 (0.919-1.325)	0.291
rs6691170[9]	1q41	T	0.34	1.06 (1.03-1.09)	rs11579490	0.902	T	0.37	1.003 (0.862-1.166)	0.974
rs10936599[9]	3q26	T	0.24	0.93 (0.91-0.96)	rs7621631	1	A	0.21	0.997 (0.8334-1.191)	0.97
rs16892766[8]	8q23	C	0.07	1.32 (1.21-1.44)	rs2437844	0.925	A	0.08	1.128 (0.8711-1.461)	0.36
rs6983267[8]	8q24	T	0.48	0.83 (0.79-0.87)	rs6983267	-	T	0.45	0.871 (0.751-1.009)	0.065
rs10795668[8]	10p14	A	0.33	0.91 (0.86-0.96)	rs706771	0.965	A	0.33	0.891 (0.762-1.043)	0.15
rs3802842[8]	11q23	C	0.29	1.21 (1.15-1.27)	rs3802840	1	T	0.26	1.190 (1.010-1.402)	0.037
rs11169552[9]	12q13	T	0.26	0.92 (0.90-0.95)	rs11169544	1	C	0.23	0.988 (0.831-1.175)	0.891
rs7136702[9]	12q13	T	0.35	1.06 (1.03-1.09)	rs7136702	-	A	0.37	1.143 (0.984-1.328)	0.0806
rs4444235[8]	14q22	C	0.46	1.12 (1.07-1.18)	rs11623717	0.902	G	0.45	1.013 (0.875-1.173)	0.859
rs4779584[8]	15q13	T	0.19	1.19 (1.12-1.26)	rs4779584	-	T	0.15	1.389 (1.144-1.686)	8.722E-04
rs9929218[8]	16q22	A	0.29	0.88 (0.83-0.92)	rs9925923	1	T	0.29	0.940 (0.800-1.104)	0.451
rs4939827[8]	18q21	C	0.47	0.85 (0.81-0.89)	rs7226855	1	G	0.45	0.820 (0.707-0.950)	8.204E-03
rs10411210[8]	19q13	T	0.10	0.79 (0.72-0.86)	rs7252505	0.831	A	0.13	0.902 (0.722-1.127)	0.363
rs961253[8]	20p12	A	0.36	1.13 (1.08-1.19)	rs5005940	1	T	0.34	0.1078 (0.923-1.255)	0.349
rs4925386[9]	20q13	T	0.32	0.93 (0.91-0.95)	rs4925386	-	T	0.31	0.960 (0.820-1.124)	0.61

REF: reference article from which association data was taken to perform the comparison; REP: reported; CTRLS: controls; AFFY: Affymetrix

due to both false positive findings and artefacts from the calling algorithm[26], or to real differences between both populations leading to dissimilar abilities to tag the real causative variant. A PCA analysis on the EPICOLON samples compared to the WTCCC control cohorts and the data from the HapMap3 CEU and TSI populations showed clear differentiation between the Northern and Southern European populations. Although not significant, SNP loadings also evidenced principal component 3 to be exclusively driven by a region of chromosome 8 (7.2-12Mb) where a common inversion is known to occur[27,28], whereas Eigenvectors 4-7 were driven by HLA-A locus in the 6q21.2-21.3 region of chromosome 6, which has been also described as highly variable between populations[29]. Given this evidence, we compared the MAFs in the 64 EPICOLON SNPs with those in the WTCCC cohorts and HapMap3 populations and detected discrepancies in the frequencies of 10 of these markers at 8 genomic loci. Therefore, we proceeded on to replicate these SNPs in an independent Spanish cohort.

One of the SNPs, rs11987193, was favourably replicated in both this second stage and the pooled analysis. The T allele of this marker appears to have a protective effect over CRC risk. The rs11987193 SNP is located in the 8p12 locus, 128kb downstream *DUSP4*. This gene is a member of the dual kinase phosphatase family, which are well-known tumour supressors[30]. They act through the downregulation of MAP kinases, thus preventing cellular proliferation and differentiation. Deletions in this gene have already been described to happen in other types of cancers, such as those of the breast[31] and lung[32]. In the case of CRC, *DUSP4* expression appears to be modulated by *KRAS* mutations[33].

The fact that this SNP was not replicated during the initial assessment of the association signals in the CORGI cohort, together with the evident MAF discrepancies, could be a sign of differences in the tagging of the real causative variant amongst populations. Even when Europeans are presumed to be genetically homogeneous, it is not unrealistic to believe that punctual LD variations may be actually happening within populations, and that these may constitute a certain impediment in our ability to replicate association signals. Further evaluation of this marker in other Southern European populations with similar MAFs is encouraged before analysing the relationship between this variant and CRC risk susceptibility in Northern cohorts.

Our GWAS study has succeeded in the replication of 6 of the 14 already-described loci. Given the population-specific differences seen so far, we consider this an important achievement, since most of these association signals had not been previously evaluated in Southern European cohorts.
Outstandingly, we have accomplished the identification of a new CRC risk variant at 8p12, determined by rs11987193. The peculiarities of this locus may have important repercussions on subsequent analysis. For this reason, the eventual identification of the real variant is of uttermost importance. Finer mapping of the locus, coupled with additional replication efforts in larger cohorts will be needed to fully ascertain the relationship between this variant and disease in other populations.

Acknowledgments

We are sincerely grateful to all patients participating in this study who were recruited in 11 Spanish hospitals as part of the EPICOLON project. This work was supported by grants from Fondo de Investigación Sanitatia/FEDER (08/1276, 08/0024, PS09/02368), Xunta de Galicia (PGIDIT07PXIB9101209PR), Ministerio de Ciencia e Innovación (SAF 07-64873) and Fundación Privada Olga Torres (CRP). We acknowledge the Spanish National DNA Bank (BNADN) for the availability of the samples.

This study makes use of data generated by the Wellcome Trust Case-Control Consortium. A full list of the investigators who contributed to the generation of the data is available from www.wtccc.org.uk. funding for the project was provided by the Wellcome Trust under award 076113 and 085475.

References

1. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78-85.
2. de la Chapelle A. Genetic predisposition to colorectal cancer. Nat Rev Cancer 2004;4:769-80.
3. Tomlinson I, Webb E, Carvajal-Carmona L, et al. A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. Nat Genet 2007;39:984-8.
4. Tomlinson IP, Webb E, Carvajal-Carmona L, et al. A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. Nat Genet 2008.
5. Tenesa A, Farrington SM, Prendergast JG, et al. Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. Nat Genet 2008.
6. Jaeger E, Webb E, Howarth K, et al. Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. Nat Genet 2008;40:26-8.
7. Broderick P, Carvajal-Carmona L, Pittman AM, et al. A genome-wide association study shows that common alleles of SMAD7 influence colorectal cancer risk. Nat Genet 2007;39:1315-7.
8. Houlston RS, Webb E, Broderick P, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40:1426-35.
9. Houlston RS, Cheadle J, Dobbins SE, 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;42:973-7.
10. Pinol V, Castells A, Andreu M, et al. Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis

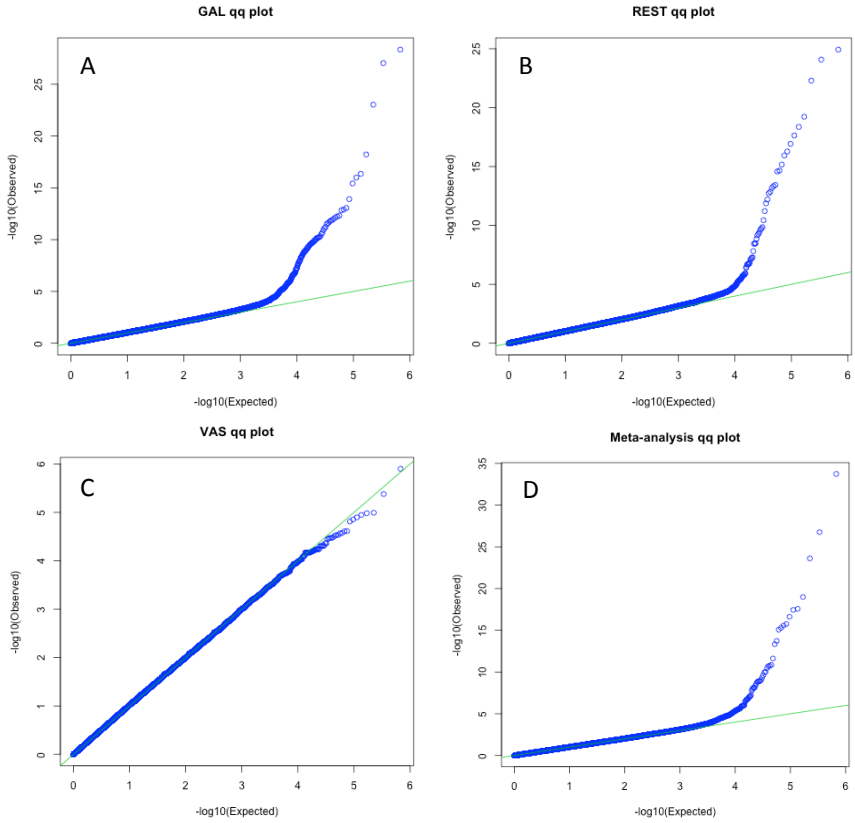
colorectal cancer. JAMA 2005;293:1986-94.
11. Fernández-Rozadilla C, de Castro L, Clofent J, et al. Single nucleotide polymorphisms in the Wnt and BMP pathways and colorectal cancer risk in a Spanish cohort. PLoS One 2010;5(9).
12. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008;40:1253-60.
13. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.
14. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2:e190.
15. Price AL, Weale ME, Patterson N, et al. Long-range LD can confound genome scans in admixed populations. Am J Hum Genet 2008;83:132,5; author reply 135-9.
16. International HapMap Consortium, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851-61.
17. Morris JA, Randall JC, Maller JB, Barrett JC. Evoker: a visualization tool for genotype intensity data. Bioinformatics 2010;26:1786.
18. Schwarzer R. Meta-analysis programs. Behavior Research Methods 1988;20:338-.
19. Petitti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine. Oxford University Press, USA, 2000.
20. Higgins J, Thompson SG. Quantifying heterogeneity in a meta‐analysis. Stat Med 2002;21:1539-58.
21. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet 2010;11:499-511.
22. 1000 Genomes Project Consortium, Durbin RM, Abecasis GR, et al. A map of human genome variation from population-scale sequencing. Nature 2010;467:1061-73.
23. de Bakker PI, Ferreira MA, Jia X, Neale BM, Raychaudhuri S, Voight BF. Practical aspects of imputation-driven meta-analysis of genome-wide association studies. Hum Mol Genet 2008;17:R122-8.
24. Johnson AD, Handsaker RE, Pulit SL, Nizzari MM, O'Donnell CJ, de Bakker PIW. SNAP: a web-based tool for identification and annotation of proxy SNPs using HapMap. Bioinformatics 2008;24:2938.
25. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 2007;447:661-78.
26. Miclaus K, Chierici M, Lambert C, et al. Variability in GWAS analysis: the impact of genotype calling algorithm inconsistencies. The pharmacogenomics journal 2010;10:324-35.
27. Iafrate AJ, Feuk L, Rivera MN, et al. Detection of large-scale variation in the human genome. Nat Genet 2004;36:949-51.
28. Kidd JM, Cooper GM, Donahue WF, et al. Mapping and sequencing of structural variation from eight human genomes. Nature 2008;453:56-64.
29. Sachidanandam R, Weissman D, Schmidt SC et al. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature 2001; 15;409(6822):928-33.
30. Keyse SM. Dual-specificity MAP kinase phosphatases (MKPs) and cancer. Cancer Metastasis Rev 2008 ;27(2):253-61.
31. Armes JE, Hammet F, de Silva M et al. Candidate tumor-suppressor genes on chromosome arm 8p in early-onset and high-grade breast cancers. Oncogene 2004;23(33):5697-702.
32. Chitale D, Gong Y, Taylor BS et al. An integrated genomic analysis of lung cancer reveals loss of DUSP4 in EGFR-mutant tumors. Oncogene 2009;28(31):2773-83.
33. Gaedcke J, Grade M, Jung K et al. Mutated KRAS results in overexpression of DUSP4, a MAP-kinase phosphatase, and SMYD3, a histone methyltransferase, in rectal carcinomas. Genes Chromosomes Cancer 2010;49(11):1024-34.

Supplementary Material

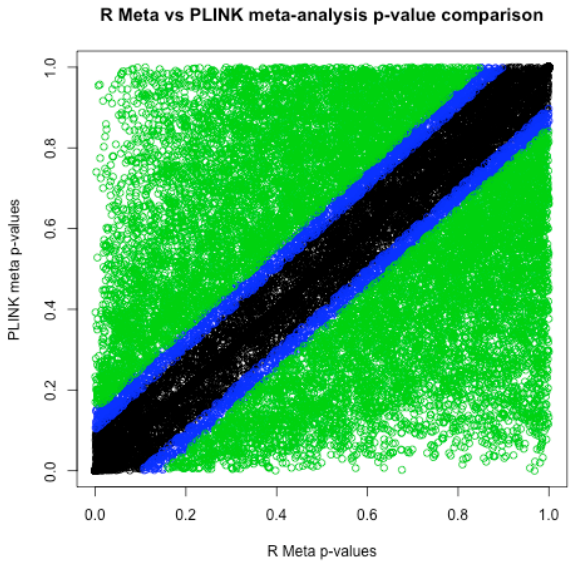
Supplementary Table 1: Phase I and Phase II cohorts. Main features and sample distribution of the stages. Gender count, hospital of collection and age statistics for cases and controls are shown for each phase.

	GENDER (MALE/FEMALE)	AGE MEDIAN (95% CI)	HOSPITAL OF COLLECTION/COHORT (number of samples)
PHASE I: 881 CASES	550/332	71.2 (70.5-71.9)	Hospital Universitari Trias i Pujol (35)
			Hospital del Mar (123)
			Hospital Clinic (91)
			Hospital General Universitario de alicante (46)
			Hospital Donostia (97)
			Hospital Universitari Arnau de Vilanova (44)
			Hospital Sant Pau (157)
			Hospital do Meixoeiro (214)
			Hospital de Calahorra (15)
			Hospital Royo Villanova (22)
			Hospital Universitario Central de Asturias (37)
			Spanish National DNA bank (0)
PHASE I: 667 CONTROLS	392/275	65.7 (64.7-66.7)	Hospital Universitari Trias i Pujol (20)
			Hospital del Mar (73)
			Hospital Clinic (0)
			Hospital General Universitario de alicante (12)
			Hospital Donostia (70)
			Hospital Universitari Arnau de Vilanova (33)
			Hospital Sant Pau (89)
			Hospital do Meixoeiro (175)
			Hospital de Calahorra (1)
			Hospital Royo Villanova (0)
			Hospital Universitario Central de Asturias (0)
			Spanish National DNA bank (194)
PHASE II: 1436 CASES	875/561	69.6 (69.0-72.2)	Hospital Gregorio Marañón (104)
			Hospital Sant Pau (125)
			Catalan Institute of Oncology (439)
			Complexo Hospitalario Universitario de Santiago (153)
			EPICOLON I (510)
			Spanish National DNA bank (105)
PHASE II: 1780 CONTROLS	1068/712	52 (51.4-52.7)	EPICOLON I (450)
			Spanish National DNA bank (1330)

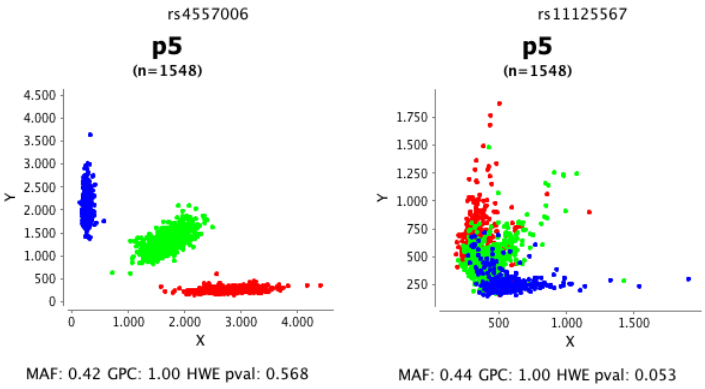
Supplementary Figure 1. Q-Q plots of p-value distribution. A: GAL; B: REST; C: VAS; D: meta-analysis.



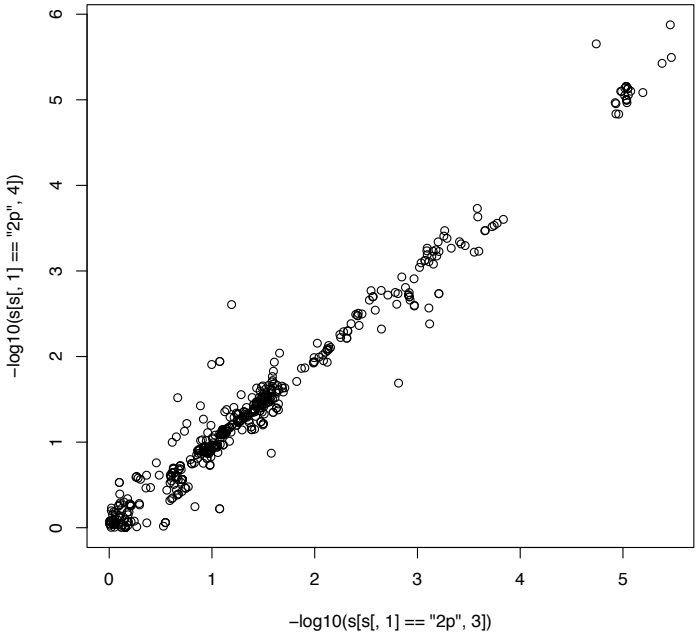
Supplementary Figure 2. P-value comparison of R META vs PLINK meta-analyses. Black: differences<0.1; blue: differences (0.1-0.15); green: differences >0.15



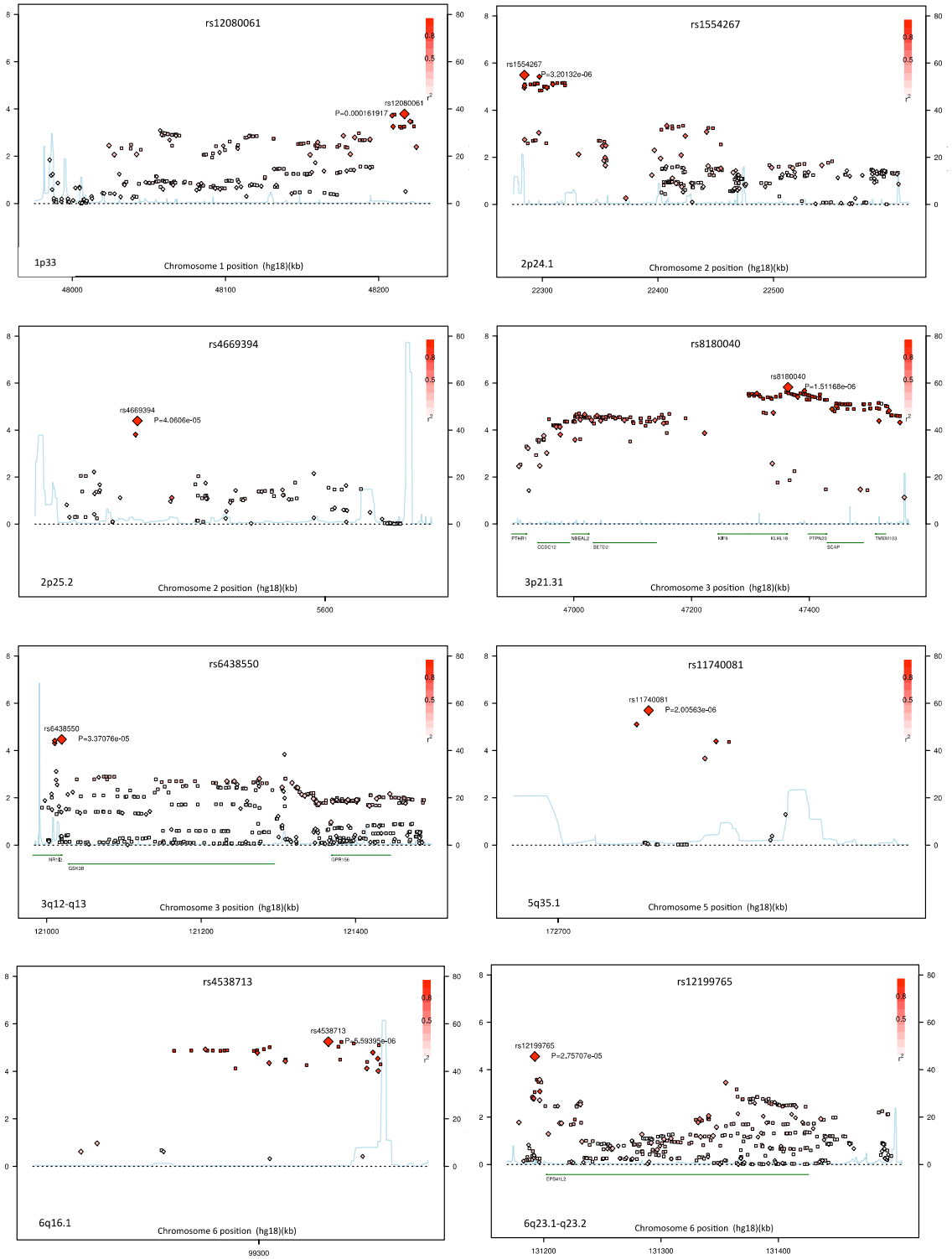
Supplementary Figure 3. Evoker plots. Left: successful calling; genotype clusters match intensity ones; right: calling error; genotypes are wrongfully called in some batches, and thus they don´t match intensity clusters.



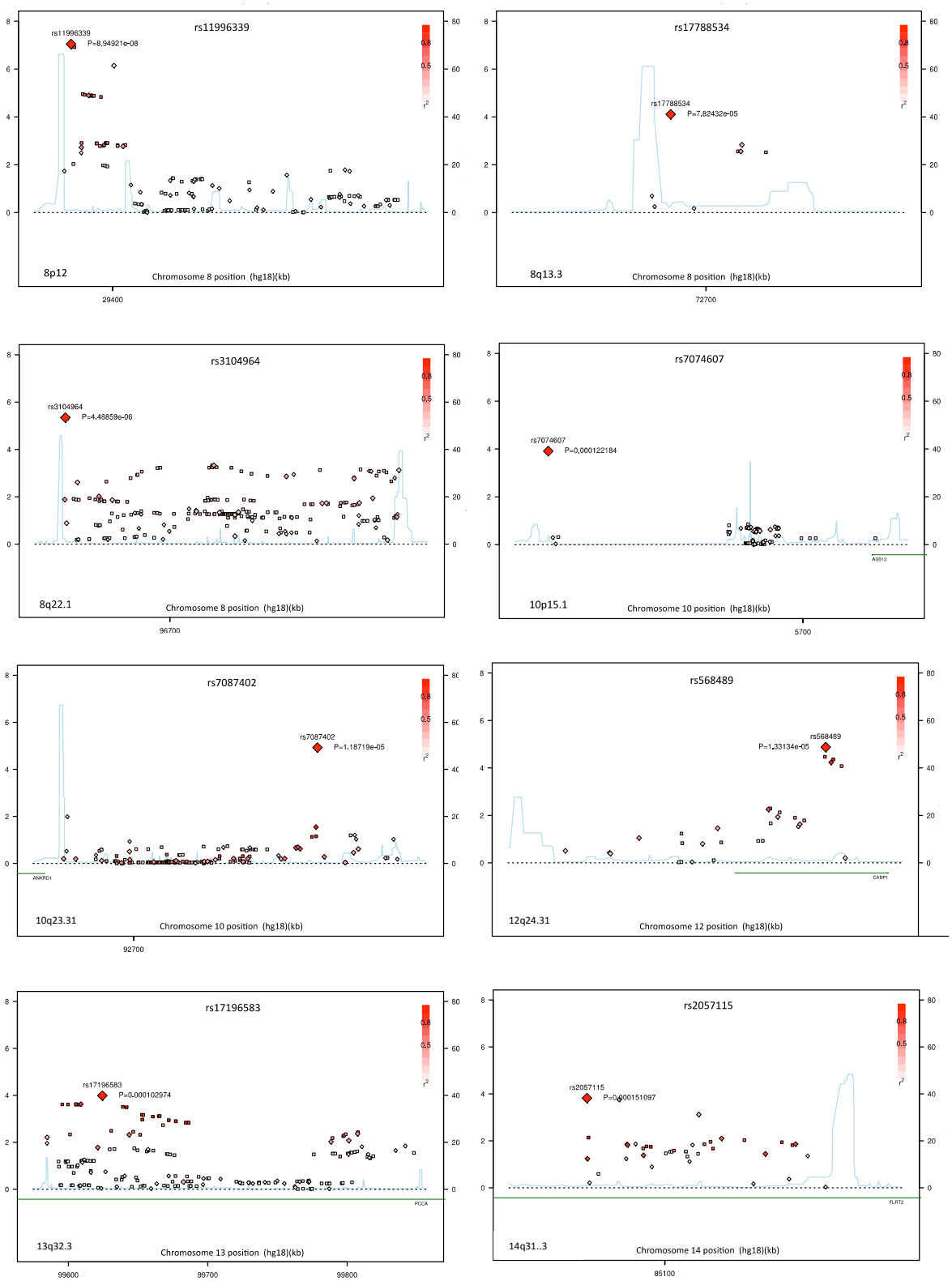
Supplementary Figure 4. Correlation plot for p-values in the two imputation strategies. An example from locus 2p25.2 is shown on the good correlation values obtained in imputation by the two different approaches.



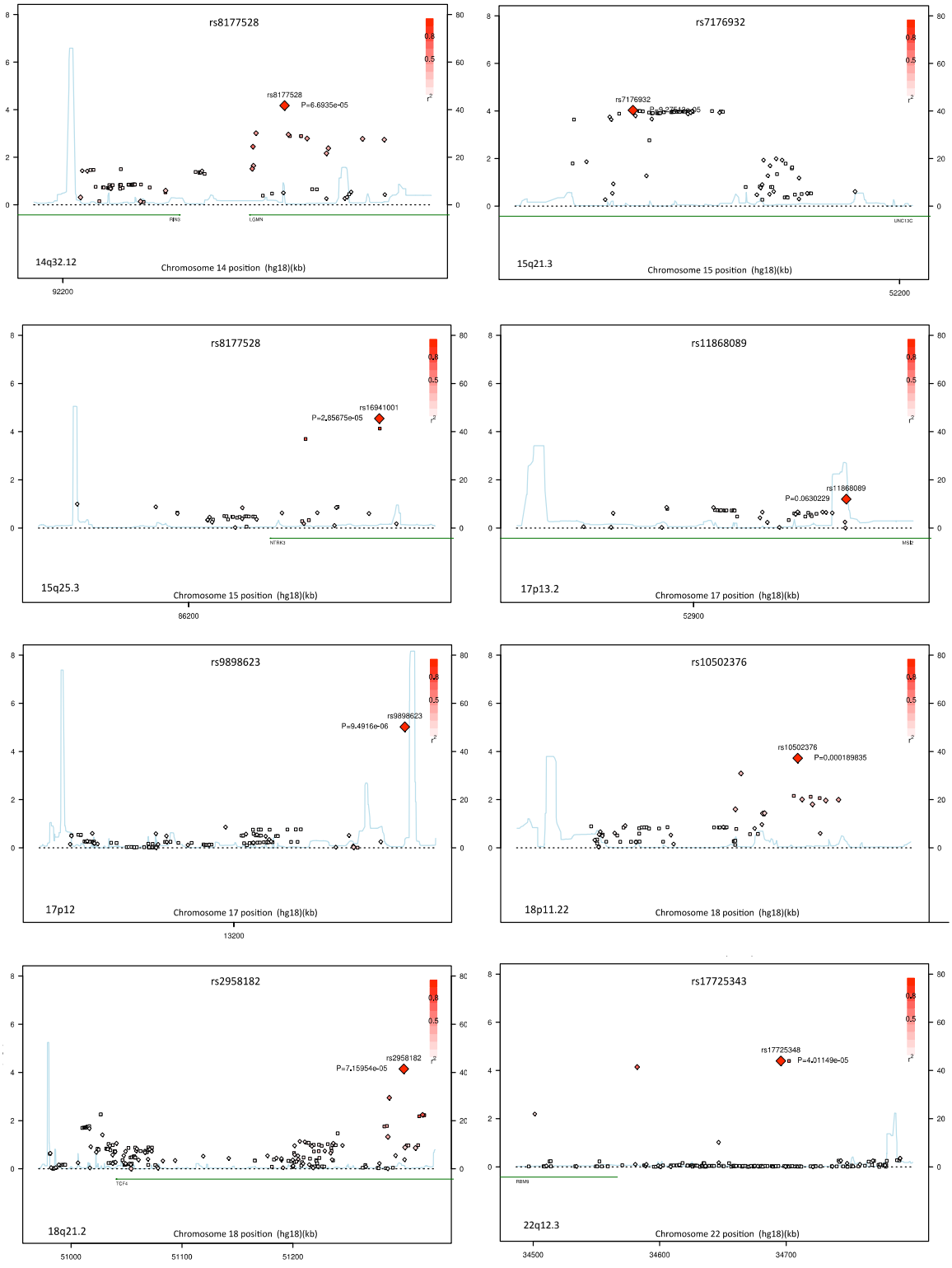
Supplementary Figure 5. Imputation plots for the 24 loci associated with CRC in EPICOLON. P-value plots for the imputed markers in the associated regions. Diamonds represent typed SNPs, squares depict imputed markers; the biggest diamond is the best-associated SNP in the region, irrespective of typed/imputed status; red grading represents LD relationships. X axis: Chromosome location; Y axis: observed (-logP); Z axis: Recombination rate (cM/Mb).



Supplementary Figure 5. (Continuation I).



Supplementary Figure 5. (Continuation II).



Chapter 4:
A Genome-Wide Association Study on Copy-Number Variation and Colorectal Cancer risk

Manuscript in preparation

A Genome-Wide Association Study on Copy-Number Variation and Colorectal Cancer risk

Fernandez-Rozadilla C[1], Cazier JB[2], Tomlinson I[2], Brea-Fernández A[1], Bujanda L[3], Bessa X[4,] Andreu M[4], Jover R[5], Llor X[6], Castells A[7], Castellví-Bel S[7], Carracedo A[1], Ruiz-Ponte C[1] for the EPICOLON consortium

[1]Galician Public Fundation of Genomic Medicine (FPGMX)-Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER)-Genomics Medicine Group-Hospital Clínico Santiago de Compostela-University of Santiago de Compostela, Spain,; [2]Wellcome Trust Centre for Human Genetics, Univseristy of Oxford, UK; [3]Colorectal Cancer Multidisciplinary Unit, Donostia Hospital, University of the Basque Country, San Sebastián, Spain; [4]Gastroenterology Department, Hospital del Mar, Barcelona, Spain; [5]Gastroenterology Department, Hospital General de Alicante, Alicante, Spain; [6]Section of Digestive Diseases and Nutrition, University of Illinois at Chicago, Chicago, IL, USA; [7]Department of Gastroenterology, Hospital Clínic, CIBERehd, IDIBAPS, University of Barcelona, Barcelona, Spain; for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association

Colorectal cancer (CRC) is a complex disease, and therefore its development is determined by the combination of both environmental factors and genetic variants. Although genome-wide association studies (GWAS) of SNP variation have conveniently identified 14 susceptibility loci, a significant proportion of the observed heritability is yet to be explained. Common copy-number variants (CNVs) are one the most important genomic sources of variability, and hence a potential source of variation to explain part of this missing genetic fraction. We have performed a GWAS on CNVs in 881 cases and 667 controls from a Spanish cohort, to explore the relationship between common structural variation and CRC development. Eleven of the common CNVs analysed in our study showed considerable potential to represent susceptibility variants. Nevertheless, we recommend additional characterisation of these CNVs by independent methods, as well as replication in larger cohorts in order to unequivocally ascertain the relationship between these variants and CRC.

Keywords: GWAS, CNV, colorectal cancer, susceptibility variant, tagSNP, Birdseye, QuantiSNP

Introduction

Colorectal cancer (CRC) is one of the most important forms of malignancy in the world, accounting for almost 50,000 deaths every year[1]. According to the Common Disease-Common Variant hypothesis, the architecture of CRC inherited predisposition is thought to be mainly explained by a combination of moderate/low-penetrance variants that interact amongst themselves to determine which individuals finally develop the disease[2]. Genome-wide association studies (GWAS) on SNPs have succeeded in the identification of 14 loci that influence the risk of CRC development at 8q24, 18q21.1, 15q13.3, 11q23.1, 8q23.3, 10p14, 14q22.2, 16q22.1, 19q13, 20p12.3[3], 1q41, 3q26.2, 12q13.13 and 20q13.33[4]. However, these altogether can only explain around 7% of the excess heritability observed, and hence, it is believed that many other variants are yet to be discovered that could account for the missing heritability proportion.

Although GWAS have been a decisive tool for the discovery of new risk variants for common diseases, they have mainly relied on the evaluation of genomic variability in the form of SNPs. Even when these markers are by far the most abundant forms of polymorphisms in the genome[5], other forms of genetic variation such as CNVs, also account for a high proportion of human polymorphic sequences[6]. Several studies have already highlighted the importance of CNVs and their potential implication in the genetic susceptibility to common diseases[7,8].

It is commonly believed that most high-frequency CNVs are well tagged by common SNPs[9,10]. Nonetheless, the presence of CNVs in a given region often triggers experimental difficulties in the determination of SNP genotypes. For instance, it has been described that sequence deletions usually result in Mendelian errors, and duplications may result in both deviations from Hardy-Weinberg equilibrium or increases in

missing genotype rates[11]. Hence, such locations have purposefully been excluded from the design of SNP genotyping arrays, resulting in an underrepresentation of CNV regions. Surveys on the distribution and features of CNVs throughout the genome have thus been consistently biased, and wider, more comprehensive scans may be important in the examination of the relationship between CNV changes and susceptibility to common diseases. Fortunately, the latest generation of genotyping assays has specifically accounted for this problem, and specific CNV probes are now implemented to attain a wider coverage in their genomic distribution.

Consequently, we decided to explore the possibility of CNV variation playing a part in CRC susceptibility by carrying out a GWA study in a Spanish cohort. For this, we used an array (Affymetrix 6.0) that allows for specific both SNP genotyping as well as the simultaneous targeting of copy-number (CN) variable regions across the genome[12]. Given the lack of consensus on a standard analysis procedure, we have chosen to use two different CNV calling algorithms, Birdsuite´s Birdseye[12] and QuantiSNP[13], to reduce the chances of false positive findings.

Materials and Methods

Study populations. Total number of individuals was 881 for CRC cases and 677 for controls. All cases and 473 controls belonged to the EPICOLON II Project. Details on the project and samples have been described elsewhere[14,15]. An additional 194 controls were obtained from the Spanish National DNA bank (www.bancoadn.org). DNA was extracted from frozen peripheral blood by standard procedures in mixed case-control batches. All samples were obtained with written informed consent, and reviewed by the ethical board of the corresponding hospital, in accordance with the tenets of the Declaration of Helsinki.

CNV genotyping, calling and QC. Samples were genotyped with the Affymetrix 6.0 array, which offers coverage for around 1 million CN variable regions. CNV calling was performed with two different algorithms: *Birdseye* (the CNV-discovery component of Birdsuite[12] and QuantiSNP v2[13]. Population stratification issues were addressed by performing Principal Component Analysis with the EIGENSOFT tool *smartpca*[16] on a set of 100,000 neutral independent SNPs (maximum pairwise r-squared=0.1). Outliers (taken as samples with >5x deviations from the cluster centroid) were removed from subsequent analyses. There were no evidences of case-control differences denoting cohort substructure for any of the first 10 components. Other potentially confounding variables were also checked for as sources for stratification (Fernández-Rozadilla *et al.*, manuscript in preparation). The final sample set consisted of 1477 samples (848 cases and 629 controls).

Further QC procedures were performed to ensure the reliability of the measures: restrictions by chromosome (only autosomes were used throughout the study), filtering by each algorithm´s quality scores (LOD ≥10 in *Birdseye* and a Bayes Factor (BF) ≥30 for QuantiSNP) and checking of B allele frequency (BAF) and log-R ratio (LRR) plots. Creation of additional plots comparing several variables (MAF, scores, CNV size, copy-number state) was made in R to test the performance of both algorithms separately. It also aided in the identification of outlier CN events, with variants over 2Mb in size or a probe count>10,000 being removed from subsequent analyses.

Statistical analysis. A statistical tool, *CNVAssoc*, was developed to perform the association tests between case and control groups after copy-number had been stated at each location with the two different algorithms. The software compares the frequencies of CNV events amongst cases and controls by considering several copy-number states to be possible at each location: homozygous deletions (0), single deletions (1), three, four and five-copy status (3,4,5). Then, it counts the incidence of every one of these states against the absence of the same, creating two-by-two contingency tables that can be used to calculate Fisher´s exact test for association. It also implements a copy-number polymorphism test approach, in which each CNV is considered to behave as a common variant. By these means, the inheritance of such variation should follow a 3-state Mendelian pattern and the 3 alleles should be in HWE.

Results

The use of both Birdseye and QuantiSNP as calling algorithms allowed for the successful identification of a total of 619,199 and 453,443 copy-number changes, respectively. These were distributed in 11,331 and 5,984 CN variable

regions, or CNVRs[11]. CNVRs are described as segments defined by the overlap of CNVs detected across samples. Of these all, 1,744 and 1,243 were polymorphic at a frequency of >5% in the control population (CNVR polymorphisms, or CNPs). Counts by algorithm and CN status are shown on Table 1.

A number of plots were created to compare the performance of both algorithms. A correlation was observed in either of the cases between CNV size and frequency (Supplementary Figure 1). As expected, larger structural rearrangements over 3Mb were rare events present in only a couple of samples, whereas CNPs tended to encompass smaller regions. The size of the detected changes and quality measures (Lod Scores and Bayes Factors) seemed also to be consistent across CN states (Supplementary Figures 2 and 3). Larger CN changes seemed to appear at higher frequencies in CN events involving gain and loss of a single copy (CNs 1 and 3). This would make sense in a biological context, where larger DNA segments would have higher chances of affecting a region of the genome where loss of both alleles is unviable and a double gain compromises dosage-dependent processes. Moreover, quality score values seemed to be dependent on the number of probes and not to the size of the detected CNV (Supplementary Figures 4 and 5). This positive correlation could be explained by the incidence of a CN change over several consecutive probes diminishing the probabilities of the detection being due to signal noise.

A marked hospital batch effect was observed during the quality control procedure that split the population in three separate clouds: samples from the Donostia hospital (VAS dataset; 97 cases and 70 controls), samples from the Meixoeiro hospital (GAL dataset; 194 cases and 172 controls) and all other samples (REST dataset; 557 cases and 387 controls) (Figure 1). The existence of this population substructure and the unavailability of association methods that allowed for the correction of this phenomenon entailed the division of the cohort into three separate subgroups from this moment of the analysis on.

Before performing the association, we considered the evaluation of the amount of existing overlap between algorithms. For this purpose, we plotted the relationship between chromosomal location and CNV frequency in the control population for the largest of the subgroups (the REST dataset) (Supplementary Figure 6). Since we aimed to assess the relationship between common CNV variation and CRC susceptibility, we generated a list of all CNPs detected by either algorithm. A considerable amount of overlap was seen between methods, with fifty-six shared locations (Supplementary Table 1).

The relationship of these CNPs with CRC susceptibility was evaluated by running CNVAssoc in the REST population (Table 2). Eleven CNPs in ten loci were associated (minimum p-value in any of the segments <0.05) in the analysis of both Birdseye and QuantiSNP calls in either the same copy number or one in the same direction (loss or gain). Association in these 11 was also checked for concordance of signals in the GAL and VAS populations (Table 3).

The potential tagging of these CNPs through SNP markers, as described by previous studies on genomic CN variation is on Table 4.

Table 1. CNV counts by calling algorithm. Total counts of detected CNVs for Birdseye and QuantiSNP.

	Birdseye	CN0	CN1	CN3	CN4	QuantiSNP	CN0	CN1	CN3	CN4
Total CN changes detected	619,199	38,112 (6%)	201,478 (33%)	272,908 (44%)	106,701 (17%)	453,443	29,220 (7%)	132,944 (29%)	213,059 (47%)	76,112 (17%)
CNVR	11,331	1,114	5,723	4,555	2,398	5,984	973	3,036	3,035	1,176
CNP	1,774	325	1,065	610	496	1,243	222	433	723	232

CNVR: copy-number variable region, defined by the overlap of CNVs across samples; CNP: polymorphic CNVR, as CNP: with frequencies over 5% in the control population.

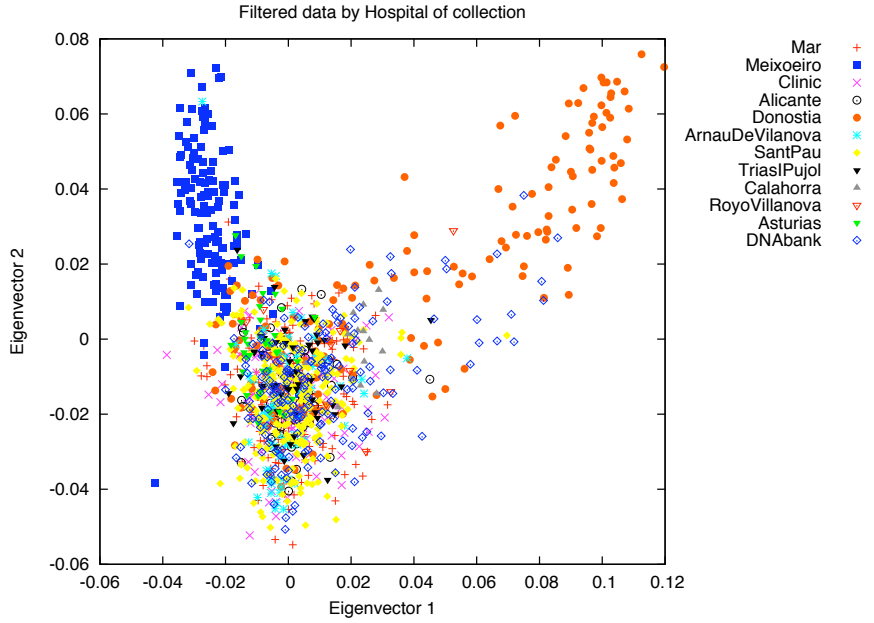


Figure 1. PCA on hospital effect based on 100.000 independent SNPs. There is a marked effect that divides the cohort into 3 main clouds: those determined by the Meixoeiro and Donosti hospitals.

Table 2. Features for the 11 associated CNPs in the REST population for both Birdseye and QuantiSNP. Association values for each algorithm are shown, as well as a description taken from the UCSC browser.

CHR	LOCATION	BS FREQ CTRLS	BS P	QS FREQ CTRLS	QS P	CN	UCSC browser
2p22.3	34,552,818 34,590,667	0.15-0.5	1.2E-05	0.10-0.17	9.1E-11	1	Serum leptin concentration QTL2; Osteoarthritis QTL2
4p16.1	9,823,257 9,844,366	0.17-0.34	0.012	0.05-0.28	4.4E-07	0,1	*SLC2A9* glucose transporter; DGV indel
4p16.1	10,001,452 10,009,766	0.05-0.26	0.01	0.05	0.008	0,1	*SLC2A9* glucose transporter; DGV indel
6q14.1	77,496,586 77,509,523	0.08-0.34	0.05	0.09	0.038	0	Osteoarthritis QTLs 16 and 22
11q11	55,130,595 55,210,152	0.08-0.41	0.0026	0.05	0.043	1	Olfactory receptor *OR4A15*; DGV deletion; Blood Pressure QTL31
15p11.1-q11.1	18,506,373 20,089,383	0.0621762	3.2E-06	0.05-0.28	0.0017	3	Centromeric
15q13.3	32,487,975 32,618,236	0.06-0.15	0.049	0.11	0.0047	0,1	Cholinergic receptor *CHRNA7* (susceptibility locus for juvenile myoclonic epilepsy); *DKFZp434L187*; DGV indel; Segmental Duplications
16p11.2-p11.1	34,324,072 34,614,572	0.13-0.20	0.013	0.07-0.16	0.044	3	*UBE2MP1*; Segmental Duplications; DGV indel; Rheumatoid arthritis QTL25; Blood pressure QTL27
17q12	36,671,885 36,684,057	0.14	2.5E-07	0.05-0.29	0.017	0,1	*ARHGAP23* and *KIAA1501*; DGV indel; Segmental Duplications; Blood Pressure QTLs 16 and 34, COPD QTLs 12 and 22; Rheumatoid arthritis QTL28
17q21.31	41,521,621 42,120,174	0.05-0.3	0.0029	0.074	0.044	3,4	Several genes; Related to Sclerosterosis, Van Buchem disease and N-acetylglutamate synthase deficiency. Downstream *BRCA1*
18q12.2	36,514,418 36,519,387	0.09-0.17	0.048	0.079	0.0023	1	COPD QTL28; Body weight QTL67

BS FREQ CTRLS: CNV frequency in controls for Birdseye analysis; QS FREQ CTRLS: CNV frequency in controls for QuantiSNP analysis;BS P: Birdseye p-value; QS P: QuantiSNP p-value; CN: copy-number; DGV: Database of Genomic Variants; COPD: chronic obstructive pulmonary disease.

Table 3. Features for the 11 associated CNPs in the GAL and VAS populations for both Birdseye and QuantiSNP. The best p-values of association are shown for each algorithm and subpopulation.

CHR	BS GAL BEST P	FREQ GAL CONTROLS	QS GAL BEST P	FREQ GAL CONTROLS	BS VAS BEST P	FREQ VAS CONTROLS	QS VAS BEST P	FREQ VAS CONTROLS
2p22.3	6.2E-05	0.28	NS	0.06-0.17	0.021	0.21	NS	0.05-0.15
4p16.1	0.029	0.31-0.47	0.00024	0.07-0.28	0.016	0.20-0.59	0.025	0.07-0.18
4p16.1	NS	0.05-0.22	NA	NA	NS	0.08-0.21	NA	NA
6q14.1	NS	0.09	NA	NA	NA	NA	NA	NA
11q11	NS	0.11-0.31	NA	NA	NS	0.17-0.37	NA	NA
15p11.1-q11.1	6.4E-10	0.11-0.15	1.1E-09	0.06-0.30	NA	NA	NS	0.07-0.24
15q13.3	0.0025	0.08-0.18	0.0094	0.09-0.13	NS	0.11-0.20	NS	0.05-0.1
16p11.2-p11.1	NS	0.10-0.14	0.039	0.05-0.13	NS	0.07-0.10	NS	0.05
17q12	NA	NA	NS	0.20	NS	0.13-0.42	0.0091	0.05-0.13
17q21.31	0.00039	0.05-0.20	0.024	0.10-0.15	0.0087	0.05-0.33	NS	0.07-0.1
18q12.2	0.031	0.16	NA	NA	NS	0.21	NS	0.05

NA: CNP was not present at the determined thresholds in the population; NS: association results were not significant at p-value<0.05.

Table 4. CNP tagging of the eleven CNPs by SNPs. The best tag-SNPs by platform and also in HapMap are depicted, as well as their pairwise properties with the corresponding CNP. Data was obtained through the data made available by the Wellcome-Trust Case-Control Consortium (WTCCC)[10].

CHR	WTCCC LOCATION	MAF	BEST AFFY6 TAG	R^2	BEST ILLUMINA TAG	R^2	BEST HAPMAP TAG	R^2
2p22.3	34,548,934-34,590,089	0.372	rs12104507	0.534	rs10179790	0.989	rs10495822	0.926
4p16.1	9,820,419-9,843,644	NA	rs6826450	0.954	rs9990501	0.943	rs231	0.955
4p16.1	10,001,049-10,012,579	0.253	rs4302457	0.994	rs4302456	0.994	rs4302456	0.995
6q14.1	77,495,977-77,517,068	0.269	rs9447790	0.592	rs9447791	0.961	rs9447791	0.973
11q11	55,202,577-55,214,079	0.26	rs654189	1	rs11230088	1	rs1944862	1
15p11.1-q11.1	18,689,010-18,894,182	0.007	rs12594870	0.022	rs12593328	0.080	rs17134298	0.071
	19,044,664-19,093,683	NA	rs6599965	0.072	rs12442343	0.061	rs7402254	0.085
	19,074,202-19,094,178	NA	rs12594870	0.080	rs12442343	0.060	rs7402254	0.102
	19,806,304-19,928,954	NA	rs4983927	0.104	rs3848222	0.109	rs7402254	0.089
	19,885,002-20,097,493	0.37	rs10220883	0.062	rs28651669	0.072	rs1303908	0.069
	19,982,358-20,068,233	NA	rs10220883	0.077	rs11259870	0.086	rs1303908	0.069
	20,047,790-20,070,506	0.095	rs4983995	0.030	rs1303908	0.031	rs1303908	0.031
	20,047,790-20,097,493	0.046	rs4983995	0.016	rs28651669	0.012	rs2664997	0.019
15q13.3	32,489,309-32,494,710	NA	rs7403222	0.300	rs4924045	0.511	rs16959239	0.540
16p11.2-11.1	34,317,021-34,615,251	0.100	rs4581708	0.377	rs11861828	0.651	rs1019991	0.667
17q12	36,675,163-36,685,731	0.265	rs9898810	1	rs2191377	1	rs8064493	1
17q21.31	41,521,114-42,139,954	0.203	rs17651507	0.467	rs17651507	0.467	rs8079215	0.463
18q12.2	36,513,895-36,520,704	0.208	rs9946719	0.657	rs9951739	0.957	rs9951739	0.957

CHR: chromosome

Discussion

Copy-number variants are an important source of variability in the genome[17]. Thus, it is possible that this type of polymorphisms, as happens with SNPs, play a part in the determination of susceptibility to complex diseases, such as CRC. We have performed a GWAS on 881 Spanish CRC cases and 667 matching controls to evaluate the relationship between structural variation and CRC risk. Given the lack of consensus on the standardisation of analytic criteria to be followed when performing CNV analyses, we chose to use two different calling algorithms: Birdsuite´s Birdseye[12] and QuantiSNP v2[13] to strengthen the validity of the associations found, as well as decrease the chances of false positive findings[18].

The comparison between algorithms showed no great differences in the overall performance of the calling procedures. Other studies have proved QuantiSNP to outrank other CNV detection methods[19]. Moreover, the improvement of this algorithm to allow for its use with of Affymetrix arrays (v2 of the software) clearly shows a good implementation for this type of data. Nevertheless, our data seem to indicate that given the current thresholds in quality control criteria, Birdseye seems to offer a greater sensitivity of CNV detection compared to QuantiSNP, since the number of CNPs detected was higher. In fact, almost all CNPs detected in QuantiSNP were also present in Birdseye. The consistency in CNP sizes in the segments that were shared amongst algorithms evidenced that this difference in detection did not correspond to a lower specificity for CNV detection over intensity noise and higher false positive rates. Similar results stating the better performance of array-matched software against other algorithms have also been described[20], although we consider that in our case, behaviour of both methods was overall analogous.

Concerning the general behaviour of the CNVs detected in our cohort, we found no significant difference in the detection rates of loss and gain events although there was a slight shift in Birdseye when the analysis was restricted only to CNVR polymorphic regions. This bias has been mentioned in the literature, and may appear as a consequence of the methodologies used to describe CNV maps being classically biased against segmental duplications[21], or as an intrinsic property of the calling algorithms themselves deriving in a decreased sensitivity to accurately detect copy-number gains. Nevertheless, results like our own have also reported in other studies[22], and thus it is likely that the differences in the detection rates are a consequence of analytic limitations.

We found that a significant number of the CNPs detected were shared between algorithms. Surprisingly, even when locations seemed to match, there were many discrepancies in CN status between Birdseye and QuantiSNP calls. Although this could be due to differences in the sensitivities of both algorithms, we find that such disagreements are not an exception, since a noticeable amount of the CN variable regions present in the databases are described as both gains and losses (indels) (http://projects.tcag.ca/variation/). The fact that even databases are so diverse in assignment of CN states could also be due to the distinct methodologies that have been used for CNV discovery and mapping, although it has also been described that the mechanisms generating losses of genomic material generate a complementary gain event[23]. It is likely that progressive fine structure analysis of these locations will provide better estimations of the CN changes underlying these loci.

The evaluation of the association of CNPs with CRC susceptibility showed copy-number states and association measures to be consistent for eleven of these CNPs in ten different loci, although p-values were sometimes very modest. The association signals were further evaluated by double-checking for correspondence in the other two subcohorts (GAL and VAS). The results were diverse, with some CNVs not attaining significance, whereas for others no evidences of such CN changes were found. These discrepancies are probably due to the smaller sample sizes of the

subsets, which could directly affect the power to detect the association signals. Likewise, the quality threshold for both the Lod Scores/Bayes Factors and the inaccurate estimation of CNV frequency in controls could as well explain the absence of CN change detection. The divergence in these results also provides with a plausible explanation for the substratification effect. Poorer quality of the callings leading to higher noise in intensities could explain the separate-cloud effect detected in quality control procedures.

None of these 11 associated CNPs seemed to affect particularly relevant genes in the CRC neoplasic process, although there was an interesting candidate at the 4p16.1 region (*SLC2A9* glucose transporter). Generally, there seemed to be an overrepresentation of CNPs lying in defined QTL regions for other diseases, such as osteoarthritis or chronic obstructive pulmonary disease (COPD). This could reflect both a shared component in the heritability of these treats or, most likely, an overrepresentation of samples suffering from these features in either our case or control groups. Interestingly, the CNP at 2p22.3 had also been investigated in our pharmacogenomic study (see chapter 5), although the association of this locus with susceptibility to 5-FU-induced nausea and vomiting could not be replicated in second-stage analysis.

Tagging of CNPs by SNPs has extensively been studied[10,24,25]. It is believed that most structural variation has somehow been indirectly assayed by GWAS performed on SNPs. This does not appear to be the case for these ten loci. Most of them are poorly tagged by either the SNP counterpart of our array, other genotyping arrays used in CRC GWA studies (mainly Illumina chips), or even HapMap[26]. This reinforces the idea that CNV variation must be inspected on its own to evaluate its implication in susceptibility to developing CRC, and possibly other disorders.

By performing a GWAS on CNV polymorphisms, we have successfully identified 11 CNPs in 10 different loci as potential candidates for CRC susceptibility loci. Notwithstanding, we consider that even when the detection by two different algorithms gives an extra reliability to the results found, the association between these 11 loci and CRC must be carefully interpreted. Replication is the most important experimental tool for assessing the validity of observed associations[27]. Besides, the assignment of CN status constitutes a problem in itself. It is recommended that confirmation of the copy-number changes in these regions is validated by independent methods, such as MLPA or qPCR, and that the resultant genotypes be in high concordance with those in our study before any additional replication in other cohorts is made.

Additionally, although the initial sample size of the study was close to the widely used 1000 cases and 1000 controls in first-stage association studies, the presence of population substratification results troublesome. Although it does not seem to increase the false-positive rate of the findings[28], it does decrease the effective sample size on which to perform the association analysis. If we were to assume that common CN variation behaves in the same way as the susceptibility SNPs described so far[3,4], then our study would be clearly underpowered to detect even the greatest of the described effects. Thus, it is likely that much larger and homogeneous cohorts are needed to detect the effects of any potential CNVs in the risk of developing CRC.

Acknowledgments

We are sincerely grateful to all patients participating in this study who were recruited in 11 Spanish hospitals as part of the EPICOLON project. This work was supported by grants from Fondo de Investigación Sanitatia/FEDER (08/1276, 08/0024, PS09/02368), Xunta de Galicia(PGIDIT07PXIB9101209PR),Ministerio de Ciencia e Innovación (SAF 07-64873) and Fundación Privada Olga Torres (CRP). We acknowledge the Spanish National DNA Bank (BNADN) for the availability of the samples.

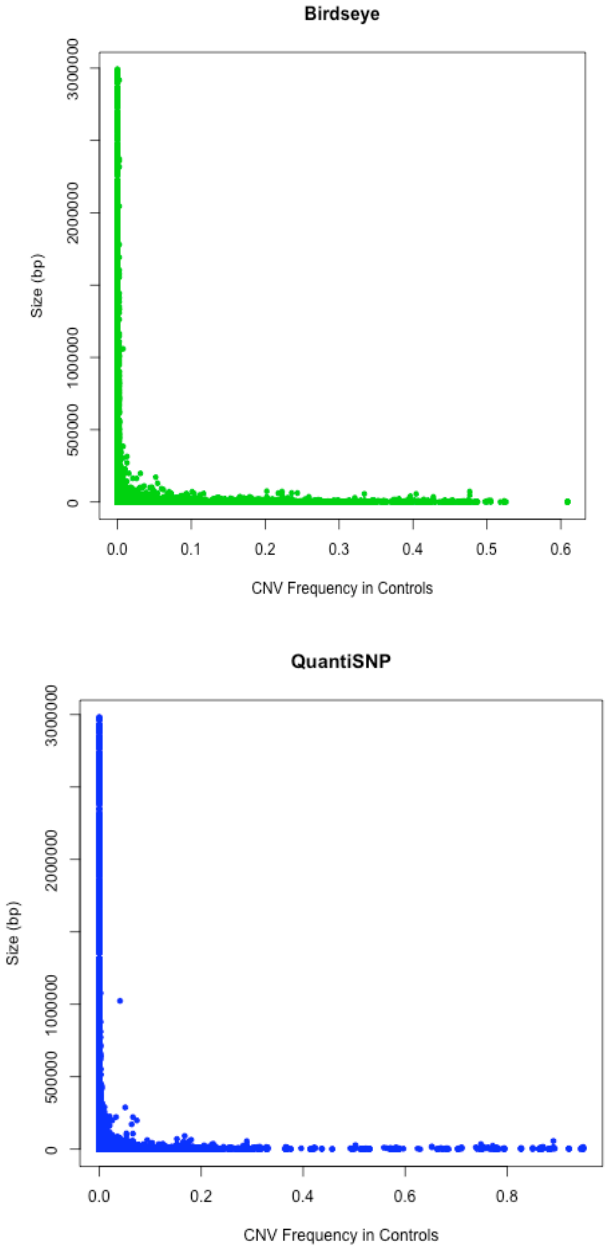
References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008:

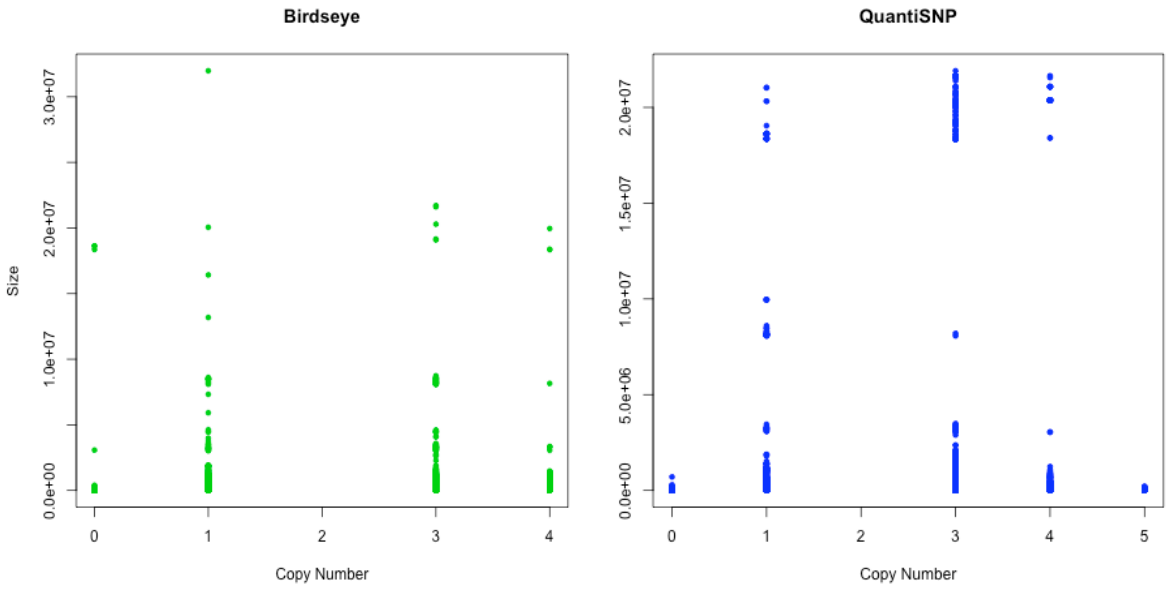
GLOBOCAN 2008. Int J Cancer 2010;127:2893-917.
2. Lander ES. The new genomics: global views of biology. Science 1996;274:536.
3. Houlston RS, Webb E, Broderick P, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40:1426-35.
4. Houlston RS, Cheadle J, Dobbins SE, 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;42:973-7.
5. Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet 2001;27:234-5.
6. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. Nature 2009; 464(7289):704-12.
7. McCarroll SA, Altshuler DM. Copy-number variation and association studies of human disease. Nat Genet 2007;39:S37-42.
8. Estivill X, Armengol L. Copy number variants and common disorders: filling the gaps and exploring complexity in genome-wide association studies. PLoS Genet 2007;3:1787-99.
9. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 2006;38:82-5.
10. Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010;464:713-20.
11. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. Nature 2006;444:444-54.
12. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008;40:1253-60.
13. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35:2013-25.
14. Pinol V, Castells A, Andreu M, et al. Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis colorectal cancer. JAMA 2005;293:1986-94.
15. Fernández-Rozadilla C, de Castro L, Clofent J, et al. Single Nucleotide Polymorphisms in the Wnt and BMP Pathways and Colorectal Cancer Risk in a Spanish Cohort. PloS one 2010;5:e12673.
16. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2:e190.
17. Feuk L, Marshall CR, Wintle RF, Scherer SW. Structural variants: changing the landscape of chromosomes and design of disease studies. Hum Mol Genet 2006;15 Spec No 1:R57-66.
18. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet 2007;16 Spec No. 2:R168-73.
19. Dellinger AE, Saw SM, Goh LK, Seielstad M, Young TL, Li YJ. Comparative analyses of seven algorithms for copy number variant identification from single nucleotide polymorphism arrays. Nucleic Acids Res 2010;.
20. Winchester L, Yau C, Ragoussis J. Comparing CNV detection methods for SNP arrays. Brief Funct Genomic Proteomic 2009;8:353-66.
21. Locke DP, Sharp AJ, McCarroll SA, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. Am J Hum Genet 2006;79:275-90.
22. Wong KK, deLeeuw RJ, Dosanjh NS, et al. A comprehensive analysis of common copy-number variations in the human genome. Am J Hum Genet 2007;80:91-104.
23. Hastings P, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. Nature Reviews Genetics 2009;10:551-64.
24. Cooper GM, Zerr T, Kidd JM, Eichler EE, Nickerson DA. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. Nat Genet 2008;40:1199-203.
25. McCarroll SA, Kuruvilla FG, Korn JM, et al. Integrated detection and population-genetic analysis of SNPs and copy number variation. Nat Genet 2008;40:1166-74.
26. International HapMap Consortium, Frazer KA, Ballinger DG, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449:851-61.
27. Scherer SW, Lee C, Birney E, et al. Challenges and standards in integrating surveys of structural variation. Nat Genet 2007;39:S7-15.
28. Wang WYS, Barratt BJ, Clayton DG, Todd JA. Genome-wide association studies: theoretical and practical concerns. Nature Reviews Genetics 2005;6:109-18.

Supplementary Material

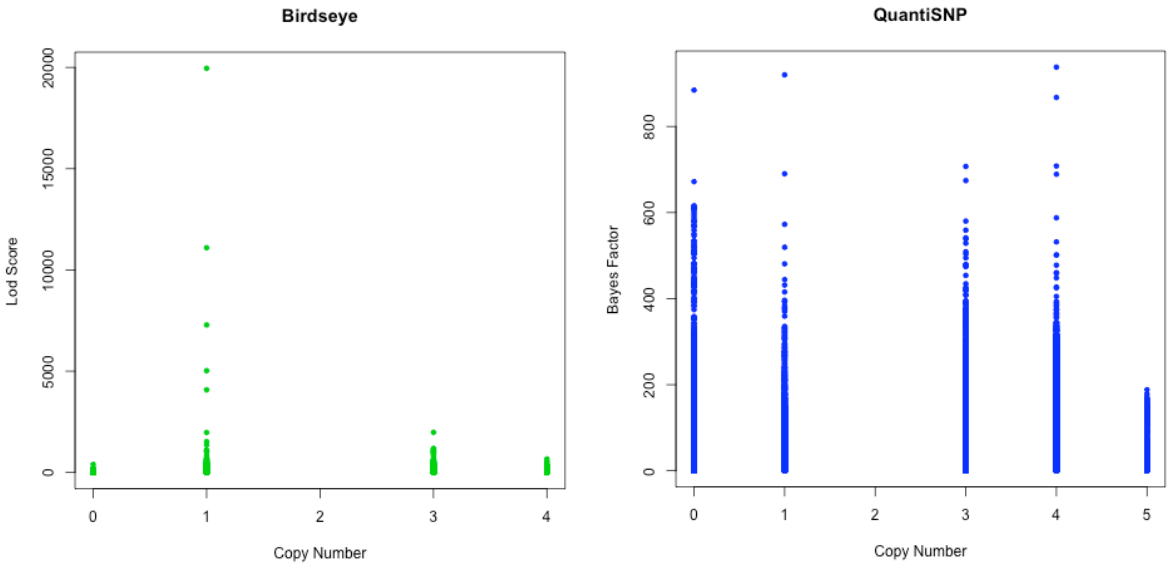
Supplementary Figure 1. CNV frequency in controls vs CNV size (bp). Larger CN rearrangements tend to be rare events, whereas frequent CN changes have sizes typically under 500kb. Top: Birdseye; bottom: QuantiSNP.



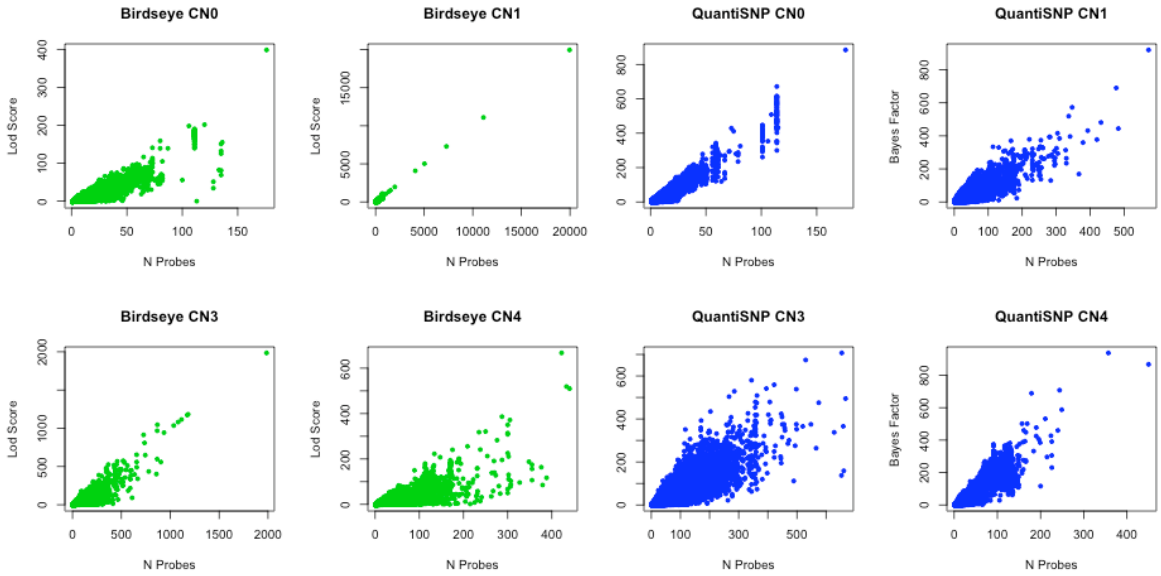
Supplementary Figure 2: CNV size versus copy-number for Birdseye and QuantiSNP. Sizes of the detected variants seem to be overall consistent for each of the algorithms between copy-number states. Left: Birdseye; right: QuantiSNP.



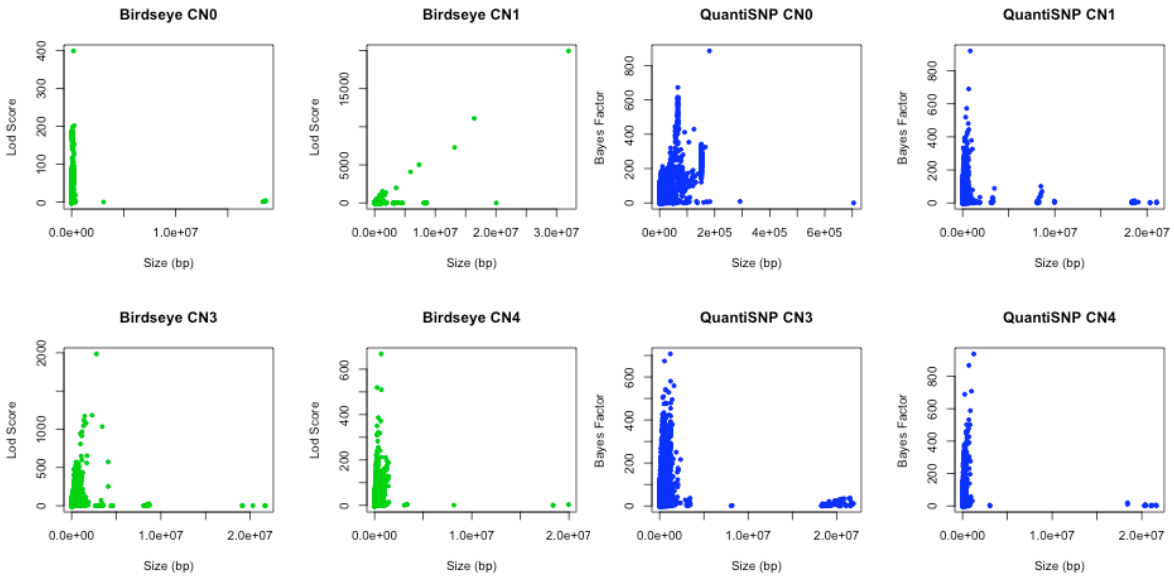
Supplementary Figure 3. Quality control scores by copy-number. Quality scores appear higher for gains/losses involving a single copy in Birdseye, whereas the distribution is homogeneous for QuantiSNP. Left: Birdseye; right: QuantiSNP.



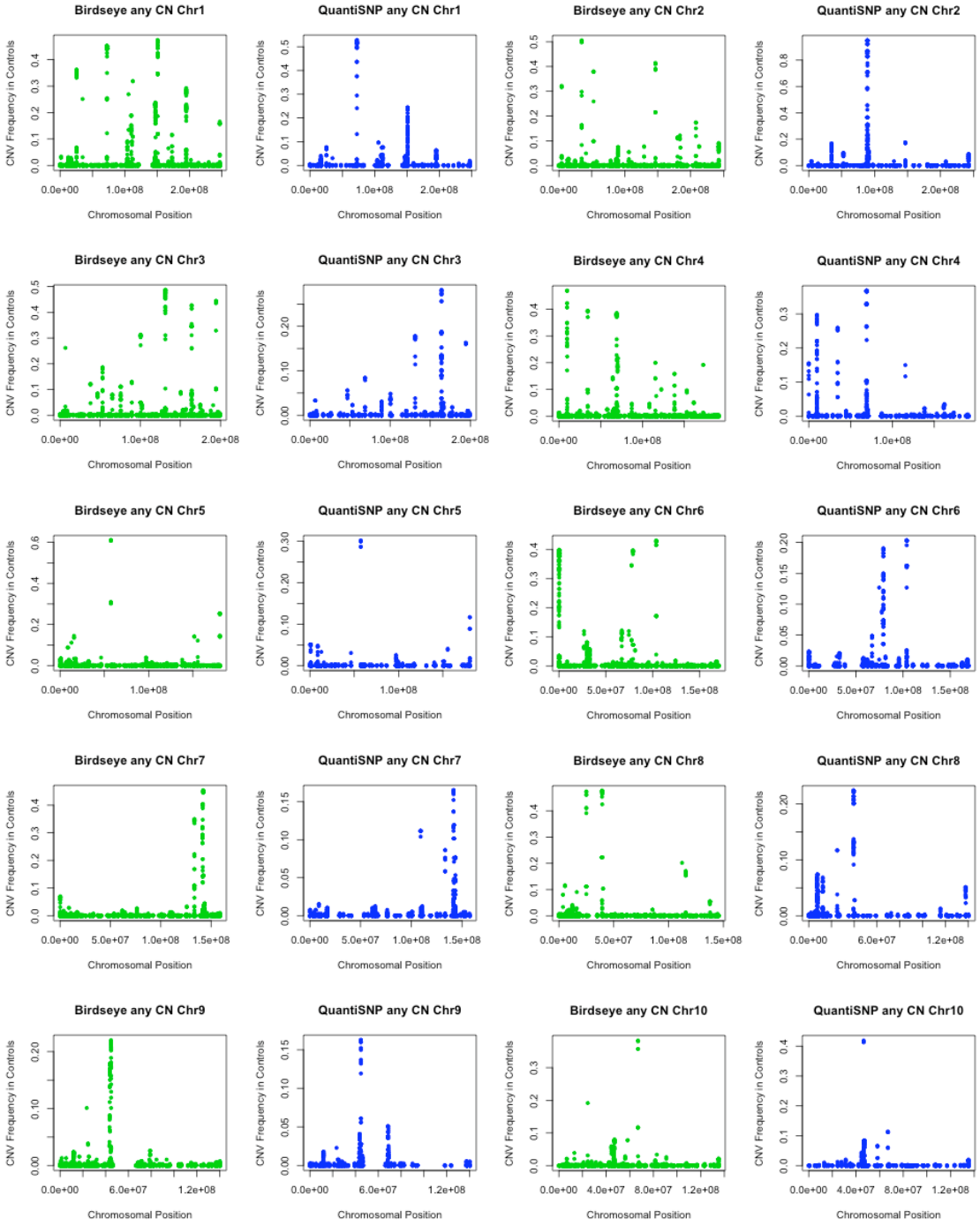
Supplementary Figure 4. Correlation between number of probes and quality score. Both quality scores seem to be dependent on number of probes. X represents number of probes and Y Lod Score (Birdseye) or Bayes Factor (QuantiSNP). Left: Birdseye; right: QuantiSNP.



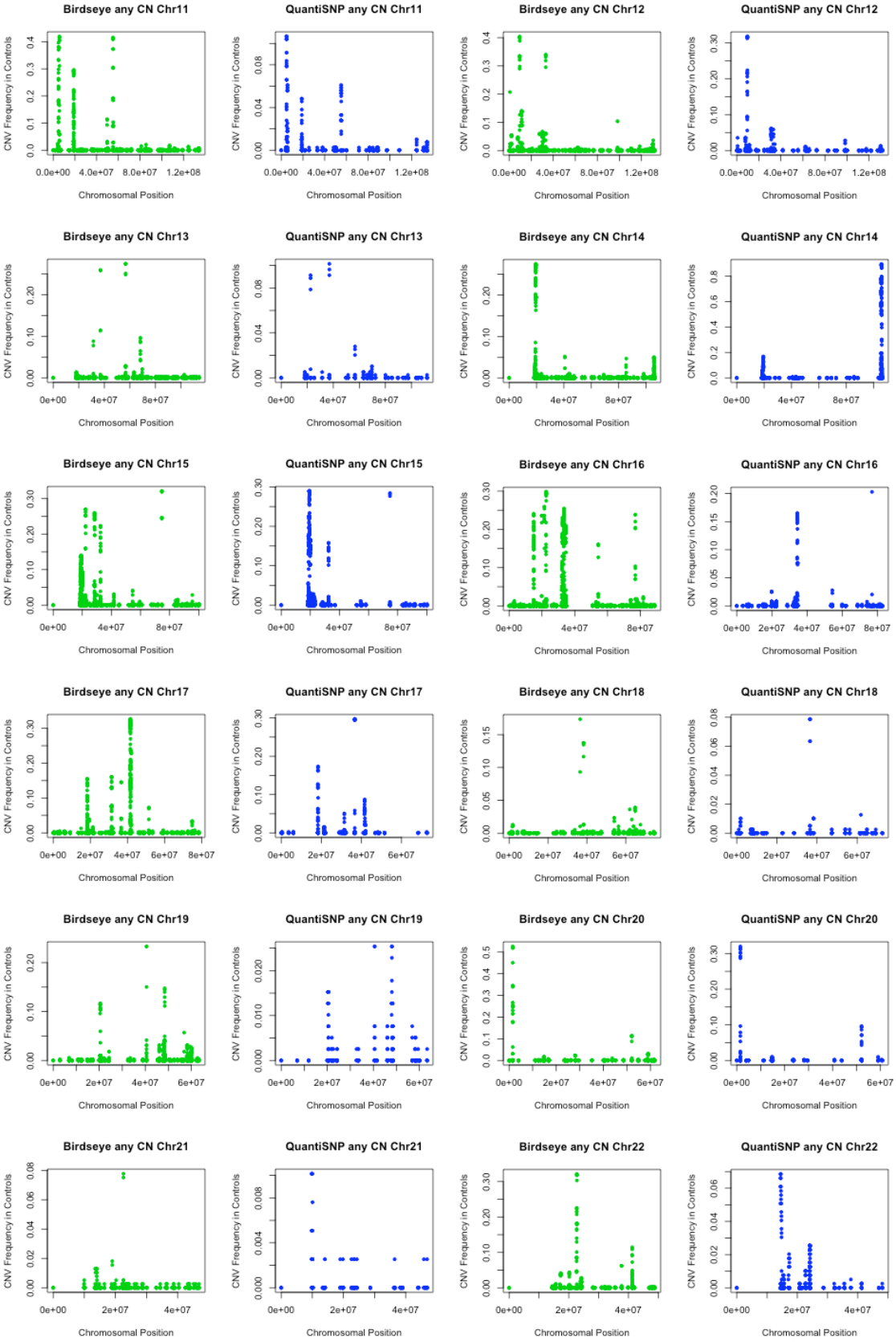
Supplementary Figure 5. Relationship between CNV size and quality. There seems to be a tendency of greater scores for larger CNVs in the case of Birdseye CN-1 variants. X represents number of probes and Y Lod Score (Birdseye) or Bayes Factor (QuantiSNP). Left: Birdseye; right: QuantiSNP.



Supplementary Figure 6. Distribution of CNVRs by chromosome. Frequencies for each CNVR are shown comparing Birdseye (green) and QuantiSNP (blue) locations in the REST control dataset; CNVRs with frequencies over 5% are CNPs; from left to right and top to bottom: chromosomes 1 to 10.



Supplementary Figure 6. (Continuation). From left to right and top to bottom: chromosomes 11 to 22.



Supplementary Table 1. Shared CNPs between Birdseye and QuantiSNP findings in the REST population. Features for the 56 CNPs that are shared between algorithms. In blue and bold, CNPs with consistent association results for both Birdseye and QuantiSNP. In pink, CNPs that have inconsistent copy number status between both algorithms.

			BIRDSEYE				QUANTISNP			
CHR	START	END	FREQ CTRLS	CN	P-VALUE	CN-P	FREQ CTRLS	CN	P-VALUE	CN-P
1	25,465,702	25,519,573	0.07-0.36	0,1	0.029	0	0.07	1	NS	-
1	72,528,701	72,583,736	0.12-0.44	0,1	0.0071	0	0.29-0.52	4	2.7E-08	1
1	150,821,799	150,853,218	0.34-0.44	0,1	NS	-	0.07-0.2	0,1	4.5E-12	0
1	151,028,534	151,035,324	0.11-0.41	0,1	NS	-	0.10-.012	0,1	6.1E-06	1
1	194,994,460	195,076,539	0.05-0.28	0,1	0,0021	1	0.06	1	NS	-
2	**34,552,818**	**34,590,667**	**0.15-0.5**	**0,1**	**1.20E-05**	**1**	**0.10-0.17**	**0,1**	**9.1E-11**	**1**
2	52,607,959	52,635,046	0.09-0.37	0,1	0.0062	1	0.08	0	NS	-
2	88,914,226	91,281,977	0.07	0,1	NS	-	/	1,3,4	2.5E-09	4
2	146,580,861	146,583,404	0.21-0.41	0,1	NS	-	0.17	1	0.0025	1
2	242,564,139	242,683,359	0.08	1	NS	-	0.07	1	NS	-
3	46,776,808	46,824,593	0.07	1	NS	-	0.06	1	NS	-
3	131,245,549	131,288,926	0.29-0.40	1	0.037	1	0.11-0.17	1	NS	-
3	163,995,338	164,109,297	0.08-0.35	0,1	0.006	1	0.08-0.27	1,3,4	2.4E-06	4
3	194,360,583	194,365,597	0.10-0.44	0,1	0.0027	0	0.16	4	7.2E-05	4
4	**9,823,257**	**9,844,353**	**0.17-0.34**	**0,1**	**0.012**	**1**	**0.05-0.28**	**0,1,4**	**4.4E-07**	**0**
4	**10,001,452**	**10,009,766**	**0.05-0.26**	**0,1**	**0.01**	**1**	**0.0532995**	**0**	**0.008**	**0**
4	34,455,242	34,501,120	0.08-0.34	0,1	NS	-	0.05-0.25	0,1	0.00039	1
4	69,043,070	69,203,991	0.10-0.38	0,1	0.007	0	0.05-0.38	1,3	3.3E-14	3
4	115,394,759	116,395,574	0.13-0.19	1	NS	-	0.15	1	NS	-
5	57,361,771	57,369,290	0.3-0.6	0,1	NS	-	0.30	4	NS	-
5	180,311,303	180,350,709	0.12-0.25	0,1	NS	-	0.08-0.11	0,1	6.1E-06	1
6	74,648,952	74,658,138	0.106218	0	NS	-	0.13	0	NS	-
6	**77,496,586**	**77,509,523**	**0.08-0.34**	**0,1**	**0.05**	**0**	**0.09**	**0**	**0.038**	**0**
6	79,025,771	79,091,904	0.08-0.34	0,1	NS	-	0.06-0.14	0,1	3.4E-10	1
6	103,844,656	103,868,754	0.17-0.42	0,1	NS	-	0.12-0.20	1,3	0.014	1
7	133,435,704	133,449,386	0.10-0.34	0,1	0.0034	1	0.08	0	NS	-
7	141,416,112	141,715,788	0.28-0.40	0,1	NS	1	0.11-0.16	1,3	3.7E-07	3
7	142,156,281	142,167,486	0.14-0.44	0,1	NS	-	0.10	0,1	NS	-
8	25,030,360	25,040,250	0.11-0.46	0,1	0.022	1	0.12	0,1	NS	-
8	39,349,340	39,506,122	0.22-0.47	0,1	NS	-	0.11-0.20	0,1	0.014	1
8	137,757,067	137,931,617	0.05	1	NS	-	0.05	1	NS	-
9	44,667,842	44,795,733	0.10-0.21	1	0.0012	1	0.06-0.16	1	NS	-
10	46,401,426	47,174,643	0.05	3	NS	-	0.05-0.41	1,3	1.1E-06	1
10	58,186,368	58,196,843	0.08	1	NS	-	0.07	1	NS	-
10	66,977,929	66,984,452	0.11-0.38	0,1	NS	-	0.05-0.11	0	NS	-
11	4,924,226	4,933,658	0.11-0.38	0,1	0.00019	1	0.06-0.1	0,1	NS	-
11	5,743,981	5,768,936	0.05-0.41	0,1	NS	-	0.06	0	NS	-
11	**55,130,595**	**55,210,152**	**0.08-0.41**	**0,1**	**0.0026**	**1**	**0.05**	**0,1**	**0.043**	**1**

CHR: Chromosome; FREQ CTRLS: CNV frequency in control population; CN-P: copy-number state of best p-value

Supplementary Table 1. (Continuation).

			BIRDSEYE				QUANTISNP			
CHR	START	END	FREQ CTRLS	CN	P-VALUE	CN-P	FREQ CTRLS	CN	P-VALUE	CN-P
12	9,521,810	9,626,952	0.08-0.40	0,1	NS	-	0.05-0.31	0,1	1.6E-11	1
12	31,171,857	31,293,957	0.06	3	NS	-	0.07	3	NS	-
12	33,191,058	33,198,641	0.10-0.33	0,1	NS	-	0.06	0	NS	-
13	36,970,023	36,982,757	0.11-0.25	0,1	NS	-	0.10	0	NS	-
14	18,801,397	19,493,212	0.06-0.25	3,4	0.00034	3	0.05-0.17	3	NS	-
15	**18,652,835**	**19,566,875**	**0.05-0.12**	**1,3**	**NS**	**-**	**0.05-0.28***	**3***	**0.0017***	**3***
15	**19,818,876**	**20,089,383**	**0.06**	**3,4**	**3.2E-06**	**3**				
15	**32,487,975**	**32,618,236**	**0.06-0.15**	**0,1,3**	**0.049**	**0**	**0.11**	**1**	**0.0047**	**1**
15	74,678,283	74,682,830	0.24-0.31	0,1	0.011	1	0.28	0	NS	-
16	**34,324,072**	**34,614,572**	**0.13-0.20**	**3**	**0.013**	**3**	**0.07-0.16**	**3**	**0.044**	**3**
16	76,929,941	76,942,266	0.06-0.23	0,1	NS	-	0.20	0	NS	-
17	18,296,117	18,415,358	0.05-0.15	0,1,4	NS	-	0.06-0.17	0,1	0.032	1
17	31,464,091	31,509,204	0.07-0.16	3,4	0,022	4	0.05	3	NS	-
17	**36,671,885**	**36,684,057**	**0.15**	**0**	**2.5E-07**	**0**	**0.05-0.29**	**1**	**0.017**	**1**
17	**41,521,621**	**42,120,174**	**0.05-0.3**	**1,3,4**	**0.0029**	**4**	**0.07**	**3**	**0.044**	**3**
18	**36,514,405**	**36,519,387**	**0.09-0.17**	**0,1**	**0.048**	**1**	**0.08**	**1**	**0.0023**	**1**
20	1,505,190	1,541,893	0.17-0.45	0,1	0.046	0	0.05-0.31	4	0.0066	4
20	52,081,215	52,092,058	0.11	1	NS	-	0.07	1	NS	-

CHR: Chromosome; FREQ CTRLS: CNV frequency in control population; CN-P: copy-number state of best p-value.
**QuantiSNP find a single CNP in this region, and thus only one value is shown.*

Chapter 5:
Pharmacogenomics in Colorectal Cancer: A Genome-Wide Association Study to predict toxicity after 5-Fluorouracil or FOLFOX administration

Submitted to The Pharmacogenomics Journal

Pharmacogenomics in Colorectal Cancer: A Genome-Wide Association Study to predict toxicity after 5-Fluorouracil or FOLFOX administration

Fernandez-Rozadilla C[1], Cazier JB[2], Crous M[3], Guinó E[3], Moreno V[3], Durán G[4], Lamas MJ[4], Paré L[5], Baiget M[5], Páez D[6], López JL[7], Cortejoso L[7], García MI[7], Bujanda L[8], González D[9], Gonzalo V[10], Rodrigo L[11], Reñé JM[12], Jover R[13], Brea-Fernández A[1], Andreu M[14], Bessa X[14], Llor X[15], Palles C[2], Tomlinson I[2], Castellví-Bel[10], Castells A[10], Carracedo A[1], Ruiz-Ponte C[1] for the EPICOLON consortium

[1]Galician Public Fundation of Genomic Medicine (FPGMX)-Centro de Investigación Biomédica en Red de Enfermedades Raras (CIBERER)-Genomics Medicine Group-Hospital Clínico Santiago de Compostela-University of Santiago de Compostela, Spain,; [2]Wellcome Trust Centre for Human Genetics, University of Oxford, UK; [3]Service of Oncology and Cancer Registry, Catalan Institute of Oncology (ICO); [4]Oncology Pharmacy Unit, Complejo Hospitalario Universitario of Santiago (CHUS), Spain; [5]Molecular Genetics Unit, Hospital de Santa Creu I Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain; [6]Medic Oncology Department, Hospital de Santa Creu I Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain; [7]Pharmacogenetics & Pharmacogenomics Laboratory, Pharmacy Unit, Hospital General Universitario Gregorio Marañón, Madrid, Spain; [8]Colorectal Cancer Multidisciplinary Unit, Donostia Hospital, University of the Basque Country, San Sebastián, Spain; [9]Gastroenterology Department Hospital de Santa Creu I Sant Pau, Universitat Autònoma de Barcelona, Barcelona, Spain; [10]Department of Gastroenterology, Hospital Clínic, CIBERehd, IDIBAPS, University of Barcelona, Barcelona, Spain; [11]Gastroenterology Department, Hospital General de Asturias, Oviedo, Spain; [12]Gastroenterology Department, Hospital Universitari Arnau de Vilanova, Lleida, Spain; [13]Gastroenterology Department, Hospital General de Alicante, Alicante, Spain; [14]Gastroenterology Department, Hospital del Mar, Barcelona, Spain; [15]Section of Digestive Diseases and Nutrition, University of Illinois at Chicago, Chicago, IL, USA, for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association.

The development of genotyping technologies has allowed for wider-coverage screenings of the hidden heritability underlying the observed variation in drug administration outcome. We have performed a Genome-Wide Association Study (GWAS) on 221 colorectal cancer (CRC) patients that had been treated with the anticancer chemotherapy agents 5-fluorouracil (5-FU) alone or in combination with oxaliplatin (FOLFOX). After evaluating a set of ~1M markers of each SNPs and copy-number variants (CNVs) in a two-stage procedure, we found that none of the CNVs but eleven of the SNPs (rs10158985 at 1q42.12, rs4128317 at 2p22.3, rs17626122 at 2q33.3, rs839533 at 2q34, rs16857540 at 3q26.31, rs10106310 at 8q21.3, rs2465403 at 8q24.12, rs10876844 at 12q13.2, rs10784749 at 12q15, rs11080058 at 17q11.2 and rs670454 at 18p11.22) showed evidences of association with adverse drug reaction (ADR) phenotypes. Ours is the first study to explore the genetic basis underlying inter-individual variation in toxicity responses to the administration of 5-FU or FOLFOX in CRC patients in a genome-wide scale. We encourage future efforts in the pharmacogenomic field, since the characterisation of such variants would help on the optimisation of the chemotherapy protocols, thereby reducing health-care costs.

Keywords: pharmacogenomics, 5-fluorouracil, FOLFOX, ADR, GWAS, colorectal cancer

Introduction

It has been known for many decades now, that there is an important inter-individual variation in an individual´s response to drug administration[1]. This variation may be represented by differences in the delivery of the drug molecule, or by factors that affect drug targeting. This divergence usually results in either the lack of the desired therapeutic effect, or the occurrence of adverse drug reactions (ADRs) with any factors such as age, sex, intake of other drugs and inheritance influencing this outcome[2].

CRC is the third most frequent form of neoplasm, and an important cause of morbidity in the developed world[3]. There has been increasing evidence from clinical trials that chemotherapy treatment greatly improves the chances of healing and survival in CRC patients with stages III or greater[4]. Five-fluorouracil has been the cornerstone for first-line CRC systematic chemotherapy treatment for many years[5], and its combination with oxaliplatin (FOLFOX) has also become a very popular treatment of choice for CRC patients[6]. However, the toxicities associated with the administration

of these drugs have sometimes overshadowed the benefits they deliver. Patients treated with 5-FU, or its oral prodrug capecitabine commonly present gastrointestinal and haematopoietic toxicities, whereas FOLFOX-treated patients are exposed to developing sensory neuropathy, which may endure even long after chemotherapy cessation[7]. All these side effects are thought to be mostly due to the narrow therapeutic indexes of most anticancer drugs.

Until recently, the investigation of the inheritance factors underlying the diverse response to CRC chemotherapy agents had mainly focused on candidate-gene studies, in which variants in genes coding for proteins involved in specific pathways, such as drug absorption, metabolism or target molecules, were screened for evidences of their association with therapy outcome. For instance, variants in candidate genes such as *DPYD*[8], *TYMS*[9] or *UGT1A1*[10] have already been linked to the development of ADRs in CRC patients treated with chemotherapy. However, these large-effect phenotypes might not apply to the majority of drugs. It is expected that for common pharmacogenetic treats, same as for most diseases, the inheritance patterns behind these responses are complex, with an additive interplay of multiple variants in the determination of the final outcome[11]. In this sense, the simultaneous study of higher numbers of variants has become increasingly necessary in order to evaluate the full contribution of inheritance to drug response. GWAS may therefore be an important tool for this purpose. The main advantage of this type of studies against gene-based strategies, is that they may be able to identify variants in genes or pathways that have, up-to-now, not been implicated in mediating drug response[12]. Nonetheless, there have still been no reports of GWAS in relation to colorectal cancer chemotherapy, neither for drug response, nor for ADRs. The discovery of the genetic factors underlying these expected heritability may be fundamental for the adjustment of therapies and/or dosage in order to achieve a better outcome.

Thus, we decided to perform an unprecedented GWAS on a cohort of samples that had either been treated with 5-FU/capecitabine or FOLFOX, with the aim to shed a light on the genetic variation behind a series of gastrointestinal (diarrhoea, mucositis, nausea/vomiting), haematological (anemia, neutropenia, leukopenia, thrombocytopenia), and neurological (oxaliplatin-related neuropathy) ADRs.

Materials and Methods

Study populations. Samples from phase I were 221 colorectal cancer patients collected through the EPICOLON II Project[13], a multicentric epidemiology overview of the prevalence and attributes of colorectal cancer in the Spanish population. All patients had received adjuvant or palliative chemotherapy, in the case of colon cancers, whereas rectal cancer patients had undergone neoadjuvant treatments. Clinical and toxicity-related information was obtained from each of these individuals using a standardarised form. Ninety-three of these patients had been dosed with 5-FU/capecitabine in monotherapy as first-line treatment, and the remaining 133 had been administered with FOLFOX. Median age for 5-FU patients was 72 with a range of (26-86), whereas average was 70.59 (68,46-72,72); age median for FOLFOX patients was 69 (range 42-85) with an average 65,85 (64,18-67,53). Gender proportions were 57,47/42,53% (male/female) for 5-FU and 67,83/32,17% for FOLFOX individuals.

Samples from phase II were 821 colorectal cancer patients collected at four different centers: Hospital Sant Pau (Barcelona), Catalan Institute of Oncology (ICO, Barcelona), Hospital Gregorio Marañón (Madrid) and Complexo Hospitalario Universitario de Santiago (Santiago de Compostela), that had undergone chemotherapeutical treatment after a colorectal cancer diagnosis. Of these, 491 had received first-line treatment with 5-FU/capecitabine and 330 with FOLFOX. Age features and sex proportions were 62,29 (60,44-64,15) average, 62 (21-83) median and 58,67/41,33% (male/female) for 5-FU, and 60,80 (58,46-63,14) average, 62 (26-75) median and a 65,93/34,07% male/female ratio for FOLFOX patients.

All samples were of Caucasian European origin and from Spain. DNA was obtained from peripheral blood by standard extraction methods.

Ethics statement. The study was approved by both each of the institutional review boards of the hospitals where samples were collected and the "Comité Ético de Investigación Clínica de Galicia". All samples were obtained with written informed consent reviewed by the ethical board of the corresponding hospital, in accordance with the Declaration of Helsinki.

Adverse effects and phenotype coding. Toxicity responses for seven different adverse reactions were recorded for each patient during their chemotherapy treatment; these were: anemia, leukopenia, thrombocytopenia, neutropenia, nausea/vomiting, diarrhoea and mucositis. Additionally, peripheral neuropathic responses, which are typical of FOLFOX treatments, were also considered for such patients. The severity of each variable was documented following the guidelines of the WHO-Toxicity grading scale for determining the severity of adverse effects[14]. Considering the discrepancies in the literature about the consideration of grade-1 toxicity patients as either cases or controls, we decided to take into account a number of variables for phenotype coding. Firstly, we defined the variable "delay" as a deviation of ≥2/3 weeks (depending on cycle number being below, or over 6, respectively) in the total number of weeks the treatment endured over the expected time. This *expected time* was calculated considering the usual 5-FU/capecitabine and FOLFOX administration regimes (4 weeks for 5-FU/FOLFOX and 3 for capecitabine) and the number of chemotherapy cycles received (i.e. for a patient that had received 5 cycles of capecitabine, delay was considered if the real number of weeks exceeded 17). Secondly, we also acknowledged the reductions in chemotherapeutical agent as registered in the filled toxicity forms. Given this, we considered grade-1 toxicity individuals as cases whenever: a) the number of chemotherapy cycles received was ≤4 for either 5-FU/capecitabine or FOLFOX; b) the number of cycles was >4 but the sample had suffered dosage reduction and/or had been qualified as "delayed" according to the abovementioned criteria. If neither of these conditions was fulfilled, grade-1 toxicity individuals were regarded to as controls. All other toxicity grades were considered as cases, regardless of the circumstances. A simplification of the classification algorithm is shown on Figure 1.

Besides, due to the low numbers of cases for some of the adverse effects (neutropenia and thrombocytopenia mainly) and the fact that all of the cell types involved come from the same stem cell type, we decided to consider these two, along with anemia and leukopenia altogether as a single variable (*haematological* adverse effects).

SNP genotyping and QC. The Affymetrix SNP 6.0 array was used for the screening of variants related to toxicity responses in colorectal cancer patients treated with 5-FU and oxaliplatin in stage I of the study. For phase II genotyping, Sequenom MassARRAY (Sequenom Inc. San

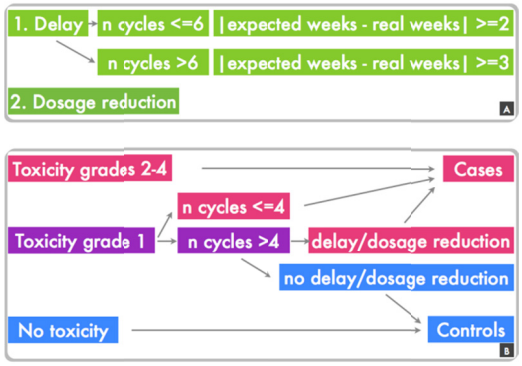


Figure 1. Simplification of the phenotype coding criteria. Figure A depicts the two variables that are considered for phenotype classification of grade 1 toxicity patients: delay and dosage reduction, whereas figure B clarifies how samples are sorted into case and control groups.

Diego, USA) was used. Genotype calling from array intensities on phase I was achieved with the use of the *Birdseed* algorithm, included in Birdsuite v1.4[15]. Later conversion of the data into PLINK v1.07[16] format was resolved using in-house scripts. By these means, we obtained an initial set of 909.622 SNPs. The minor allele frequency (MAF) filtering threshold was set to ≥0.15 instead of the usual 0.05 in order to avoid noise signalling derived from the low sample size numbers. Genotyping success rates, both for samples and markers, were set at 95%, and individuals were also checked for concordance of genders between clinical recorded data and the Affymetrix-asigned sex. Hardy-Weinberg Equilibrium (HWE) was also tested, and markers with p-values <0.001 were removed from subsequent analyses. Differencial missingness in cases *vs* controls was assessed for each phenotypic group, although no additional markers were to be removed because upon this criterion. Principal Component Analysis (PCA) on genotypes, with the aim to detect any possible underlying population stratification, was performed with the help of the EIGENSOFT *smartpca* tool[17]. No differences were observed between case and control groups for neither the 5-FU nor the FOLFOX groups and any of the ADR phenotypes for the first 10 eigenvectors (Supplementary Figure 1). After these filtering procedures, 497,366 SNPs in 88 individuals, and 497,913 SNPs in 115 individuals remained for the 5-FU and FOLFOX treatments, respectively. The total number of case/control counts for each phenotype is shown on Table 1.

We also considered performing an evaluation of the association that had previously been linked to ADR susceptibility after 5-FU and FOLFOX treatments by means of an overview on already

Table 1. Sample count for each of the phenotypes. Summary of sample sizes for 5-FU and FOLFOX groups with each ADR in phase I.

/	5-FU CASES	5-FU CTRL	FOLFOX CASES	FOLFOX CTRL
Diarrhoea	25	63	40	75
Mucositis	9	79	13	102
Nausea/ Vomiting	8	80	23	92
Haematologic	14	74	50	65
Neuropathy	-	-	47	68

CTRL: controls

published SNPs or their closest proxies on our array. The selection of the markers was performed with the help of the PharmGKB database (http://www.pharmgkb.org/), and was restricted to only those variants with MAF>5%.

Statistical analysis. SNP association analyses were performed by logistic regression in PLINK[16]. Covariate adjustment was also used to correct for gender and severity of the toxicities (grades 1-2 *vs* 3-4) during the testing. Hospital of origin of the sample was also adjusted for in phase II analyses. Odds ratios (ORs) and 95% confidence intervals (CIs) were calculated for each marker. Plots were created using Haploview[18] and R.

CNV calling and analysis. CNV probes on the array were used to screen for structural variation. Copy-number status was estimated by two different methods: the *Birdseye* algorithm, another part of Birdsuite´s framework[15], and QuantiSNP version 2[19], which includes a modification that implements the use of the CN-probe information on Affymetrix arrays. Once the calling was made, CN variants were filtered according to their MAFs (>5%) and the probability score associated with each calling algorithm. QC thresholds were set at Lod-scores >10 for Birdseye CNVs and a Bayes Factor >30 for QuantiSNP results. Association analyses were then performed with the help of *CNVAssoc*, a home-made tool that allows for the comparison of CNV event frequencies among cases and controls. Briefly, the program considers the presence of a number of different copy-number states: homozygous deletion (0), single deletion (1), three, four or five copies (3,4,5), any loss (0+1), any gain (3+4+5), and any CNV (0+1+3+4+5). Then, for each, it counts the incidence of every one of these copy-number states against the absence of the same, creating two-by-two contingency tables that can be used to calculate Fisher´s exact test for association. It also implements a copy-number polymorphism test approach, in which each CNV is considered to behave as a common variant. By these means, the inheritance of such variation should follow a 3-state Mendelian pattern and the alleles should be in HWE. If the counts do not fit this equilibrium, then the change may as well be unstable, multiallelic or a rare CNV.

Results

The results for the SNP analysis for stage I on the 88 5-FU and 115 FOLFOX patients showed very modest association p-values, with none reaching the established genome-wide significance level. This could be mostly due to the lack of power derived from the low case numbers that were available for each of the phenotypes. The quantile-quantile plots showing the distribution of the association p-values clearly reflect this effect (Supplementary Figures 2 and 3). Considering the fact that this would imply an increase in type II error (false negative rate), we decided to further genotype in a second stage the top 5 association hits for each of the treatment-phenotype pairs, with the exception of neuropathy, for which the top 10 loci were selected because of the specificity of this adverse reaction with the intake of FOLFOX (Table 2).

As for CNVs, association p-value results after Fisher´s exact test were also moderate, both for Birdseye and QuantiSNP (Supplementary Figure 4). However, there were some outstanding regions at 2p22.3 for 5-FU-nausea/vomiting, 11p11.12 and 20p13 for 5-FU-haematologic, and 11p15.4 for FOLFOX-diarrhoea. Another CNV at 5q35.3 seemed to share susceptibility to 5-FU induced haematological effects and FOLFOX-related mucositis (Table 3). These proved interesting enough for a second-stage follow-up. All of them were already documented copy-number variation sites and were well tagged by nearby SNPs[20]

Table 2. Loci selected for replication on stage two and their associated markers. Principal features and phase I association values for the 50 SNPs selected for phase II replication are shown. Five loci were selected for each phenotype, with the exception of FOLFOX-neuropathy, for which the top 10 hits were chosen.

CHR	SNP	LOCATION	MINOR ALLELE	OR (95%CI)	P-VALUE	TREATMENT-ADR
2	**rs6713755**	223746679	G	8.321 (3.067-22.580)	3.170E-05	5-FU-diarrhoea
10	**rs887339**	29545130	G	11.260 (3.831-33.110)	1.076E-05	5-FU-diarrhoea
12	**rs10876844**	54336973	T	7.295 (2.799-19.010)	4.784E-05	5-FU-diarrhoea
12	**rs10784749**	67708296	C	0.133 (0.047-0.371)	1.174E-04	5-FU-diarrhoea
20	**rs4813881**	8916642	A	0.153 (0.060-0.389)	7.854E-05	5-FU-diarrhoea
3	**rs9861291**	180274372	G	0.060 (0.015-0.241)	7.473E-05	5-FU-haematologic
5	**rs10055794**	171054516	C	11.790 (3.431-40.490)	8.932E-05	5-FU-haematologic
16	**rs7184580**	77418447	C	7.104 (2.624-19.230)	1.138E-04	5-FU-haematologic
18	**rs1943423**	55641304	C	5.881 (2.278-15.180)	2.507E-04	5-FU-haematologic
19	**rs4805974**	39248137	A	8.477 (2.828-25.410)	1.358E-04	5-FU-haematologic
2	**rs2627043**	179290782	T	8.276 (2.476-27.660)	5.973E-04	5-FU-mucositis
3	**rs16857540**	175383269	C	7.404 (2.393-22.910)	5.130E-04	5-FU-mucositis
8	**rs2465403**	120160008	G	14.140 (3.332-59.98)	3.282E-04	5-FU-mucositis
13	**rs927553**	23706704	G	22.550 (4.157-122.400)	3.049E-04	5-FU-mucositis
18	**rs670454**	8682920	C	8.577 (2.476-29.720)	6.997E-04	5-FU-mucositis
2	**rs10182133**	75710000	G	13.730 (3.099-60.810)	5.612E-04	5-FU-nausea/vomiting
4	**rs2060645**	183344417	T	15.020 (3.246-69.48)	5.268E-04	5-FU-nausea/vomiting
4	**rs6815391**	189161973	A	11.050 (2.751-44.380)	7.086E-04	5-FU-nausea/vomiting
10	**rs7094179**	14912433	T	17.960 (3.662-88.050)	3.709E-04	5-FU-nausea/vomiting
13	**rs9300811**	102834312	G	13.950 (3.162-61.550)	5.014E-04	5-FU-nausea/vomiting
2	**rs4971347**	933636	A	0.151 (0.063-0.366)	2.701E-05	FOLFOX-diarrhoea
2	**rs250944**	234661957	C	0.242 (0-126-0-466)	2.182E-05	FOLFOX-diarrhoea
11	**rs1865282**	4787228	C	4.377 (2.244-8.538)	1.489E-05	FOLFOX-diarrhoea
12	**rs1347851**	89091109	C	5.410 (2.411-12.140)	4.227E-05	FOLFOX-diarrhoea
21	**rs2282469**	34151289	C	3.968 (2.059-7.647)	3.820E-05	FOLFOX-diarrhoea
2	**rs17626122**	206182257	G	3.638 (1.937-6.834)	5.950E-05	FOLFOX-haematologic
8	**rs10106310**	91245050	C	3.901 (2.010-7.569)	5.702E-05	FOLFOX-haematologic
13	**rs7325568**	39716284	C	3.716 (1.950-7.083)	6.608E-05	FOLFOX-haematologic
14	**rs2798389**	82120903	G	4.109 (2.023-8.347)	9.292E-05	FOLFOX-haematologic
15	**rs4243761**	23833884	G	0.279 (0.150-0.517)	5.127E-05	FOLFOX-haematologic
1	**rs520227**	188835935	C	11.550 (3.410-39.100)	8.459E-05	FOLFOX-mucositis
2	**rs4128317**	29513358	C	9.681 (3.125-29.740)	7.346E-05	FOLFOX-mucositis
2	**rs839533**	212516352	T	8.247 (2.760-24.640)	1.585E-04	FOLFOX-mucositis
14	**rs17098912**	99174082	A	7.716 (2.587-23.020)	2.480E-04	FOLFOX-mucositis
19	**rs7255865**	21596808	C	7.856 (2.674-23.080)	1.778E-04	FOLFOX-mucositis
1	**rs2389972**	85963254	G	5.234 (2.266-12.090)	1.069E-04	FOLFOX-nausea/vomiting
1	**rs10158985**	224117232	A	4.100 (1.986-8.466)	1.365E-04	FOLFOX-nausea/vomiting
6	**rs851974**	152013380	C	5.068 (2.245-11.440)	9.367E-05	FOLFOX-nausea/vomiting
8	**rs2739171**	134167847	T	0.182 0.077-0.431)	1.043E-04	FOLFOX-nausea/vomiting
9	**rs724975**	120828533	A	5.688 (2.330-13.890)	1.349E-04	FOLFOX-nausea/vomiting
1	**rs1555070**	175039581	A	3.461 (1.783-6.717)	2.437E-04	FOLFOX-neuropathy
2	**rs1097701**	129863806	A	0.287 (0.149-0.553)	1.881E-04	FOLFOX-neuropathy
3	**rs447978**	121973776	C	3.776 (1.908-7.577)	1.354E-04	FOLFOX-neuropathy
5	**rs9312960**	257082	T	3.772 (1.878-7.577)	1.910E-04	FOLFOX-neuropathy
5	**rs4957020**	367346	A	0.279 (0.143-0.546)	1.909E-04	FOLFOX-neuropathy
5	**rs12188653**	101411702	C	0.155 (0.059-0.404)	1.425E-04	FOLFOX-neuropathy
5	**rs6863960**	115023680	G	3.294 (1.748-6.208)	2.269E-04	FOLFOX-neuropathy
11	**rs17718902**	17741283	G	3.007 (1.691-5.349)	1.790E-04	FOLFOX-neuropathy
11	**rs1944118**	110857242	A	0.243 (0.129-0.458)	1.268E-05	FOLFOX-neuropathy
17	**rs11080058**	23766187	T	3.272 (1.744-6.138)	2.220E-04	FOLFOX-neuropathy

Table 3. CNV hits chosen for stage II replication. Main features of every CNV and association values in stage I are shown. Tagger SNPs for all CNVs and their pairwise r^2 measures were extracted from the data made available by the Wellcome Trust Case Control Consortium[20].

CHR	LOCATION	CN-TYPE	DETECTION ALGORITHM	MAF	P-VALUE	TAGGER SNP	R^2	TREATMENT-ADR
2p22.3	34,590,561 34,590,667	Deletion	Birdsuite	0.372	6.3E-04	rs10179790	0.997	5-FU-nausea/vomiting
11p11.12	49,667,437 49,703,156	Deletion	Birdsuite	0.060	2.8E-04	rs4466833	1	5-FU-haematologic
20p13	1,530,207 1,541,893	Deletion	Birdsuite	0.220	3.3E-05	rs2209313	0.997	5-FU-haematologic
5q35.3	180,331,966 180,350,696	Deletion	QuantiSNP	0.347	5.1E-04 2.3E.04	rs2387715	0.877	5-FU-haematologic FOLFOX-mucositis
11p15.4	5,744,656 5,765,715	Deletion	Birdsuite	0.231	2.6E-04	rs10838648	1	FOLFOX-diarrhoea

Replication values of the top 50 hits for SNPs and the 4 CNV-associated SNP taggers from phase I and pooled analysis of both phases may be seen on Supplementary Table 1. Three of these association signals were proven to be consistent throughout stages: rs10876844 at 12q13.2 with diarrhoea in patients dosed with 5-FU (pooled p =0.010; OR= 6.502 (1.552-27.23)), and rs10106310 at 8q21.3 ($p=1.193x10^{-4}$; OR=1.967 (1.393-2.776)) and rs17626122 at 2q33.3 ($p=4.851x10^{-4}$; OR=1.720 (1.268-2.332) with haematologic side effects after FOLFOX exposure. Nevertheless, there were another 8 SNPs that, although not significant in the replication stage, did show significant association values in the pooled analysis of the two stages. These were rs10158985, rs4128317, rs10784749, rs839533, rs16857540, rs2465403, rs11080058 and rs670454. The features for each of these 11 associated loci on each stage are shown on Table 4.

We also evaluated the association signals for a series of SNP variants that had been linked to either 5-FU or FOLFOX-related toxicity in the literature. None of the associations could be replicated (Table 5).

Discussion

It has been long observed that there is large variation in both the effectiveness and toxicity outcomes of drug treatment, and at least part of it has been proven to be due to inheritance[24]. The definition of these genetic factors has been the purpose of pharmacogenetic studies for almost 60 years now. For most of this time, pharmacogenetics has focused on candidate-gene approaches investigating drug-metabolising enzymes, drug transporters or drug targets. The recent advances on genome knowledge, as well as the development of new genotyping technologies during the *-omics* era have however made possible the expansion of these studies to genome-wide levels[25].

With these expectations, we followed a GWA study with the purpose of detecting new variants that could help predict the toxic effects of the administration of two of the most common chemotherapeutical drugs in colorectal cancer patients: 5-fluorouracil and FOLFOX. As major toxicity remains the main limitation to adequate dosing, the ability to predict toxicity before the administration of chemotherapy could help provide individualised treatment that would likely result in an improved outcome, both for the sake of the patient and pharmacoeconomic purposes[26].

Table 4. Summary of association results for the 11 associated loci. P-values for each stages and pooled analysis, and OR for the combined data are shown.

LOCATION	SNP	P-VALUE PHASE I	P-VALUE PHASE II	P-VALUE POOLED	OR (95%CI) POOLED	TREATMENT-ADR
3q26.31	rs16857540	5.130E-04	0.102	0.020	8.426 (1.396-50.86)	5-FU-mucositis
8q24.12	rs2465403	3.280E-04	0.853	9.426E-03	23.09 (2.158-247.1)	5-FU-mucositis
12q13.2	rs10876844	4.784E-05	0.023	0.010	6.502 (1.552-27.230)	5-FU-diarrhoea
12q15	rs10784749	1.174E-04	0.621	0.017	0.164 (0.037-0.726)	5-FU-diarrhoea
18p11.22	rs670454	6.997E-04	0.605	5.084E-03	29.19 (2.756-309.3)	5-FU-mucositis
1q42.12	rs10158985	1.370E-04	0.140	6.863E-03	1.576 (1.133-2.191)	FOLFOX-nausea/vomiting
2p22.3	rs4128317	7.346E-05	0.252	5.041E-03	1.620 (1.156-2.27)	FOLFOX-mucositis
2q33.3	rs17626122	5.950E-05	0.037	4.851E-04	1.720 (1.268-2.332)	FOLFOX-haematologic
2q34	rs839533	1.585E-04	0.096	4.426E-03	1.732 (1.186-2.527)	FOLFOX-mucositis
8q21.3	rs10106310	5.702E-05	0.057	1.193E-04	1.967 (1.393-2.776)	FOLFOX-haematologic
17q11.2	rs11080058	2.220E-04	0.322	3.589E-03	1.594 (1.165-2.182)	FOLFOX-neuropathy

The only consistent association signal for 5-FU and the ADRs considered throughout the stages was that of rs10876844 with diarrhoea. This SNP lies in the long arm of chromosome twelve, 24kb and 26kb upstream of the *METTL7B* (methyl-transferase like 7B precursor) and *ITGA7* (integrin alpha 7 isoform 1 precursor) genes, respectively. Although analysis of the LD pattern in this region seemed to show that rs10876844 and the gene blocks were not well correlated (Supplementary Figure 5), this however does not rule out the possibility that rs10876844 may be capturing part of the known, or even yet unknown variation in any of these genes. In this sense, *METTL7B* could be of particular relevance, since this very gene, as well as other family members, have been proven to interact with other drugs, namely tamoxifen, both in mice and humans[27].

The rs10106310 SNP, which was associated in our cohort with haematologic toxicity outcomes in patients treated with FOLFOX, is located 80kb upstream from the Calbindin 1 (*CALB1*) gene, in a high LD block in the 8q21.3 cytoband (Supplementary Figure 6). The presence of the calbindin gene in the nearby region could provide with a feasible explanation to the relationship between this variant and the haematologic toxicity observed. Changes of calbindin expression levels have been proven to influence the apoptotic pathway in lymphocytes[28]; thus, it is reasonable to believe that variations in the cellular availability of this protein could lead to enhanced leukocyte cell death.

The last of the association signals corresponds to rs17626122, a SNP located in the intronic region of the *PARD3B* (partitioning defective 3 homolog B) gene (Supplementary Figure 7). Although initially described in homology with Par genes, which are the major effectors of cell polarity in the mouse[29], the PARD3B protein has been shown to act with SMAD3, one of the members of the TGF-ß signalling pathway. Mutations in *SMAD3* have already been linked to colorectal adenocarcinoma development in mice[30] and wound healing capabilities[31], whereas the TGF-ß pathway in itself is a very important route in cancer development[32].

Even when phase II association values were not significant, there were some other 8 SNPs that held significance in the pooled

Table 5. Association results for a revision of markers that had been linked to toxicity outcomes after 5-FU or FOLFOX treatments. Markers were selected from previous bibliography whenever described MAFs were <5%.

CHR	SNP	GENE	MAF	PROXY	R^2	P-VALUE	OR (95% CI)	TREATMENT-ADR	REF
1	rs1801159	*DPYD*	0.167	rs17116806	0.882	0.483	1.374 (0.566-3.338)	5-FU-haematologic	21
						0.276	0.398 (0.037-2.567)	5-FU-nausea/vomiting	21
	rs1801265	*DPYD*	0.150	rs4970722	1	0.216	2.313 (0.613-8.735)	5-FU-nausea/vomiting	21
3	rs1801019	*UMPS*	0.167	rs10049380	0.838	0.737	0.822 (0.261-2.588)	5-FU-haematologic	9
						0.081	0.322 (0.0903-1.152)	5-FU-diarrhoea	9
18	rs34743033	*TYMS*	0.337	NA	-	-	-	5-FU-any toxicity	9
	rs34489327	*TYMS*	0.470	NA	-	-	-	5-FU-any toxicity	9
11	rs1695	*GSTP1*	0.375	rs7952081	0.764	0.366	0.661 (0.270-1.620)	5-FU/FOLFOX-haematologic	23
						0.360	1.295 (9.745-2.252)	FOLFOX-neuropathy	23

analysis of the two stages. It has been suggested that joint analysis of phases may be more powerful in detecting true association signals than stage-based designs particularly in studies with low sample sizes[33]. This may be so because the pooling analysis allows for an improvement in power in the overall study compared to each of the stages independently. Thus, we consider these 8 variants as interesting locations and do not fully discard them as potential susceptibility loci for adverse reactions after 5-FU and FOLFOX administration.

Of these other four 5-FU-related SNPs, rs16857540 and rs2465403 fall within the intronic regions of the neuroligin1 (*NLGN1*) and the Collectin subfamily member 10 (*COLEC10*) genes, whereas rs10784749 and rs670454 lie in intergenic locations. FOLFOX-associated variants rs10158985, rs4128317 and rs839533 also reside within genes (*TMEM63A*, *ALK* and *ERBB4*, respectively), while rs11080058 appears in a genic dessert as well.

We must also bear in mind that, for some of these SNPs (particularly for 5-FU SNPs rs2465403, rs10876844 and rs670454), OR ranges were exceptionally wide, thus possibly reflecting a type I error or the inability of the study to precisely estimate the risk associated with that variant due to sample size restrictions. ORs for FOLFOX variants seem to have more consistent ranges, thus confirming the importance of sample size in the evaluation of the risk conferred by these variants. Hence, we believe further follow-up studies in larger cohorts necessary to fully ascertain the relationship between these variants and ADR susceptibility.

On the other hand, we did not succeed in the replication of any of the association signals described at other loci in previous literature. It must however be highlighted that there is a bias that results from none of them being directly genotyped in our dataset, and the fact that r-squared values were sometimes small. Hence, we would recommend these to be additionally evaluated in other cohorts.

Scientists have largely speculated on the idea that variants contributing to pharmacogenetic traits have escaped the effects of human selection, for humans have spent most of their evolutionary history unexposed to drugs. This would have likely resulted in the frequencies of these variants rising in the populations through genetic drift, and the effect sizes of the variants being higher than the discovered for GWAS of disease susceptibility[34]. Our study however, does not confirm this hypothesis, and even when the predicted odds ratios for the associated variants we find are somewhat higher than the obtained for the majority of the GWAS and susceptibility to disease signals, the prediction of the effects of these variants is moderate. This may be yet-another reason to support the idea that pharmacogenetic traits behave just as any other complex one, and therefore the

genetic heritability behind it may mostly lie in the form of moderately penetrant common variants[12]. This would reinforce GWAS screenings in large sample sets as the ideal strategy to resolve the genetic variation behind individual differences in the response to drug administration.

In this study, we present the results of the first GWAS studying of the toxicity outcome of chemotherapeutical 5-fluorouracil and FOLFOX administration in colorectal cancer patients. We have successfully identified 11 new variants related to 5-FU-induced and FOLFOX-related side effects. Even when copy-number alterations account for a high percentage of the total variation in the genome[35], we did not succeed in identifying any signals of association regarding CNVs and the ADRs in our study. We have also proposed feasible biological mechanisms by which these association signals may modulate the toxicity responses observed. Of course, as happens with many of the GWAS-related association hits, the real biological mechanisms underlying these associations are still unknown. This however, does not rule out the relevance of the discoveries and we encourage further efforts be made in this direction. The verification of the association signals reported in this study with their presumptive ADRs would ultimately need to be ascertained by replication in larger datasets. Functional essays will eventually also be required in order to clarify the potential biological mechanisms underlying these association signals.

Acknowledgments

We are sincerely grateful to all patients participating in this study. This work was supported by grants from Fondo de Investigación Sanitatia/FEDER (08/0024, 08/1276, PS09/02368), Xunta de Galicia (PGIDIT07PXIB9101209PR), Ministerio de Ciencia e Innovación (SAF 07-64873), Fundación Privada Olga Torres (CRP), and Fundación Barrie de la Maza (Programa DIANA). We acknowledge the Wellcome Trust Case Control Consortium for making available data about SNP tagging of common CNVs (http://www.wtccc.org.uk/wtcccplus_cnv/supplemental.shtml)."

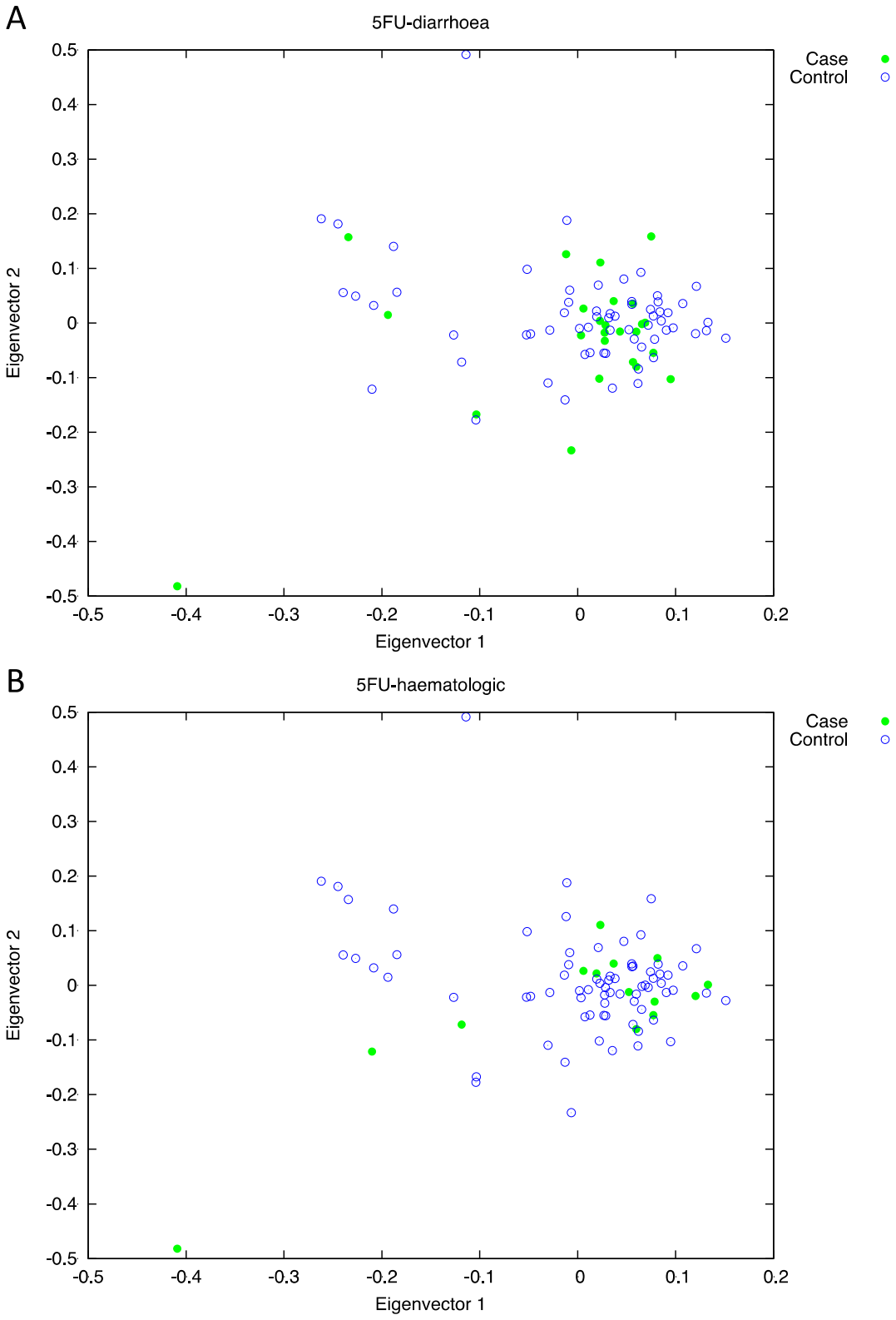
References

1. Beutler E, Dern RJ, Flanagan CL, Alving AS. The hemolytic effect of primaquine. VII. Biochemical studies of drug-sensitive erythrocytes. J Lab Clin Med 1955;45:286-95.
2. Wilke RA, Lin DW, Roden DM, et al. Identifying genetic risk factors for serious adverse drug reactions: current progress and challenges. Nature Reviews Drug Discovery 2007;6:904-16.
3. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010;127:2893-917.
4. Quasar Collaborative G, Gray R, Barnwell J, et al. Adjuvant chemotherapy versus observation in patients with colorectal cancer: a randomised study. Lancet 2007;370:2020-9.
5. Meyerhardt JA, Mayer RJ. Systemic therapy for colorectal cancer. N Engl J Med 2005;352:476-87.
6. de Gramont A, Figer A, Seymour M, et al. Leucovorin and fluorouracil with or without oxaliplatin as first-line treatment in advanced colorectal cancer. Journal of Clinical Oncology 2000;18:2938.
7. Eng C. Toxic effects and their management: daily clinical challenges in the treatment of colorectal cancer. Nat Rev Clin Oncol 2009;6:207-18.
8. Van Kuilenburg ABP, Meinsma R, Zoetekouw L, Van Gennip AH. High prevalence of the IVS14 1G> A mutation in the dihydropyrimidine dehydrogenase gene of patients with severe 5-fluorouracil-associated toxicity. Pharmacogenetics and Genomics 2002;12:555.
9. Ichikawa W, Takahashi T, Suto K, Sasaki Y, Hirayama R. Orotate phosphoribosyltransferase gene polymorphism predicts toxicity in patients treated with bolus 5-fluorouracil regimen. Clinical cancer research 2006;12:3928.
10. Iyer L, King CD, Whitington PF, et al. Genetic predisposition to the metabolism of irinotecan (CPT-11). Role of uridine diphosphate glucuronosyltransferase isoform 1A1 in the glucuronidation of its active metabolite (SN-38) in human liver microsomes. J Clin Invest 1998;101:847.
11. Meyer UA. Pharmacogenetics and adverse drug reactions. The Lancet 2000;356:1667-71.
12. Crowley JJ, Sullivan PF, McLeod HL. Pharmacogenomic genome-wide association studies: lessons learned thus far. Pharmacogenomics 2009;10:161-3.

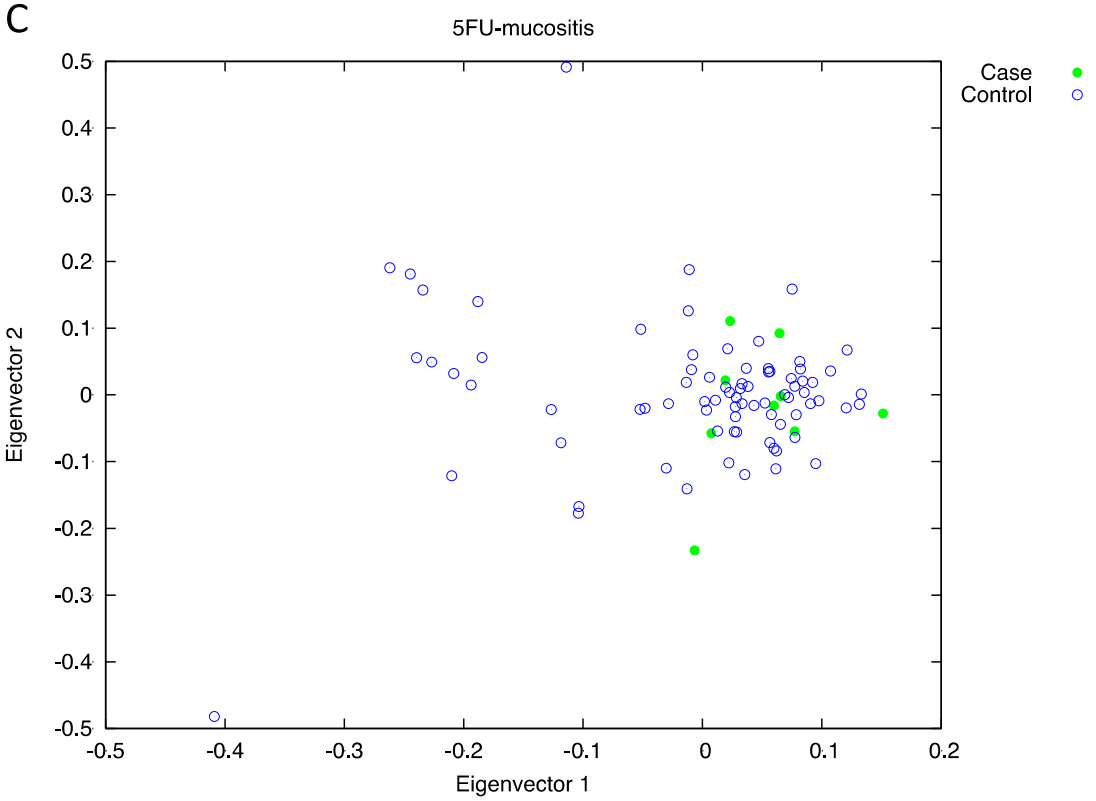
13. Pinol V, Castells A, Andreu M, et al. Accuracy of revised Bethesda guidelines, microsatellite instability, and immunohistochemistry for the identification of patients with hereditary nonpolyposis colorectal cancer. JAMA 2005;293:1986-94.
14. Trotti A, Colevas AD, Setser A, et al. CTCAE v3. 0: development of a comprehensive grading system for the adverse effects of cancer treatment* 1. 2003;13:176-181.
15. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008;40:1253-60.
16. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 2007;81:559-75.
17. Patterson N, Price AL, Reich D. Population structure and eigenanalysis. PLoS Genet 2006;2:e190.
18. Barrett JC, Fry B, Maller J et al. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 2005;21:2:263-265.
19. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35:2013-25.
20. Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010;464:713-20.
21. Zhang H, Li YM, Jin X. DPYD* 5 gene mutation contributes to the reduced DPYD enzyme activity and chemotherapeutic toxicity of 5-FU: results from genotyping study on 75 gastric carcinoma and colon carcinoma patients. Med Oncol 2007; 24:251.
22. Sohn KJ, Corxford R, Yates Z et al. Effect of the methylenetetrahydrofolate reductase C677T polymorphism on chemosensitivity of colon and breast cancer cells to 5-fluorouracil and methotrexate. J Natl Cancer Inst 2004; 96:134.
23. Lecomte T, Landi B, Beaune P et al. Glutathione S-transferase P1 polymorphism (Ile105Val) predicts cumulative neuropathy in patients receiving oxaliplatin-based chemotherapy. Clin Cancer Res 2006; 12:3050.
24. Roden DM, George Jr AL. The genetic basis of variability in drug responses. Nature Reviews Drug Discovery 2002;1:37-44.
25. Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. Nature Reviews Genetics 2003;4:937-47.
26. Weinshilboum R, Wang L. Pharmacogenomics: bench to bedside. Nature Reviews Drug Discovery 2004;3:739-48.
27. Fong CJ, Burgoon LD, Williams KJ, Forgacs AL, Zacharewski TR. Comparative temporal and dose-dependent morphological and transcriptional uterine effects elicited by tamoxifen and ethynylestradiol in immature, ovariectomized mice. BMC Genomics 2007;8:151.
28. Dowd DR, MacDonald PN, Komm B, Haussler MR, Miesfeld R. Stable expression of the calbindin-D28K complementary DNA interferes with the apoptotic pathway in lymphocytes. Molecular Endocrinology 1992;6:1843.
29. Guo S, Kemphues KJ. par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. Cell 1995;81:611-20.
30. Zhu Y, Richardson JA, Parada LF, Graff JM. Smad3 mutant mice develop metastatic colorectal cancer. Cell 1998;94:703-14.
31. Ashcroft GS, Yang X, Glick AB, et al. Mice lacking Smad3 show accelerated wound healing and an impaired local inflammatory response. Nat Cell Biol 1999;1:260-6.
32. Derynck R, Akhurst RJ, Balmain A. TGF-β signaling in tumor suppression and cancer progression. Nat Genet 2001;29:117-29.
33. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet 2006;38:209-13.
34. Cirulli ET, Goldstein DB. Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 2010;11:415-25.
35. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. Nature 2009; 464(7289):704-12.

Supplementary Material

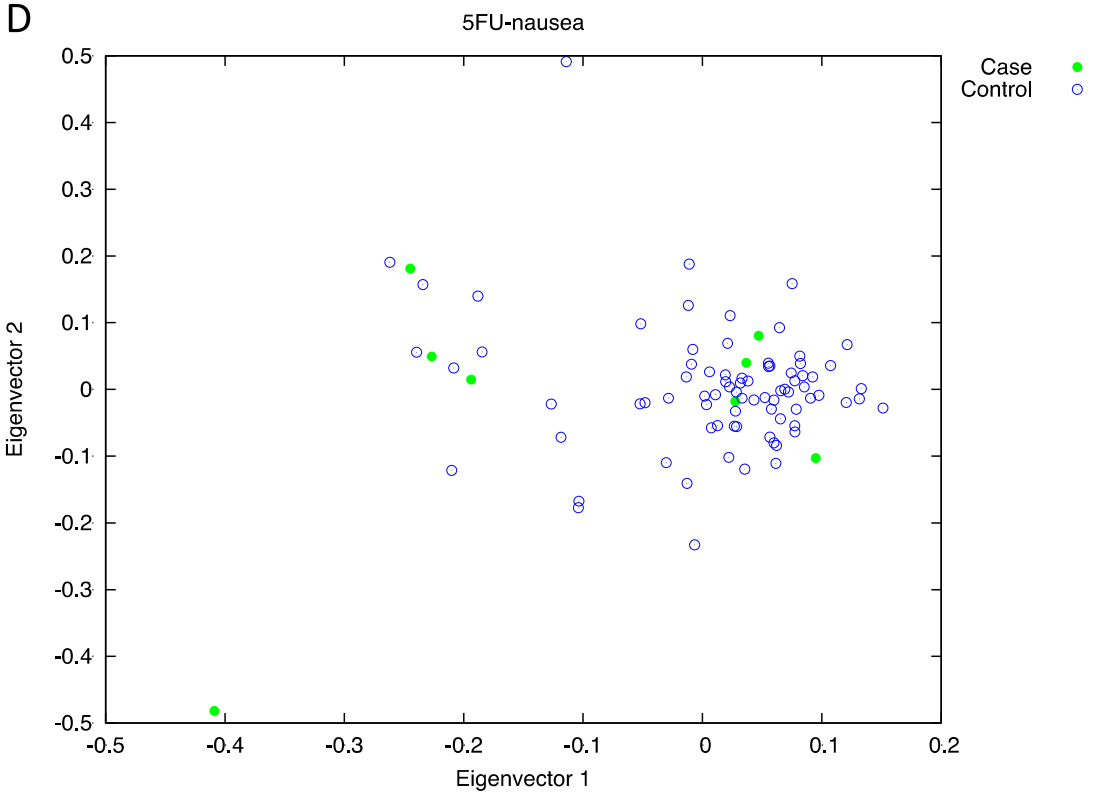
Supplementary Figure 1. PCA analysis comparing case and control groups for each of the ADR phenotypes with 5-FU and FOLFOX patients. First and second eigenvectors are shown for each treatment-ADR combination. A-D: 5-FU ADRs; E-I: FOLFOX ADRs.



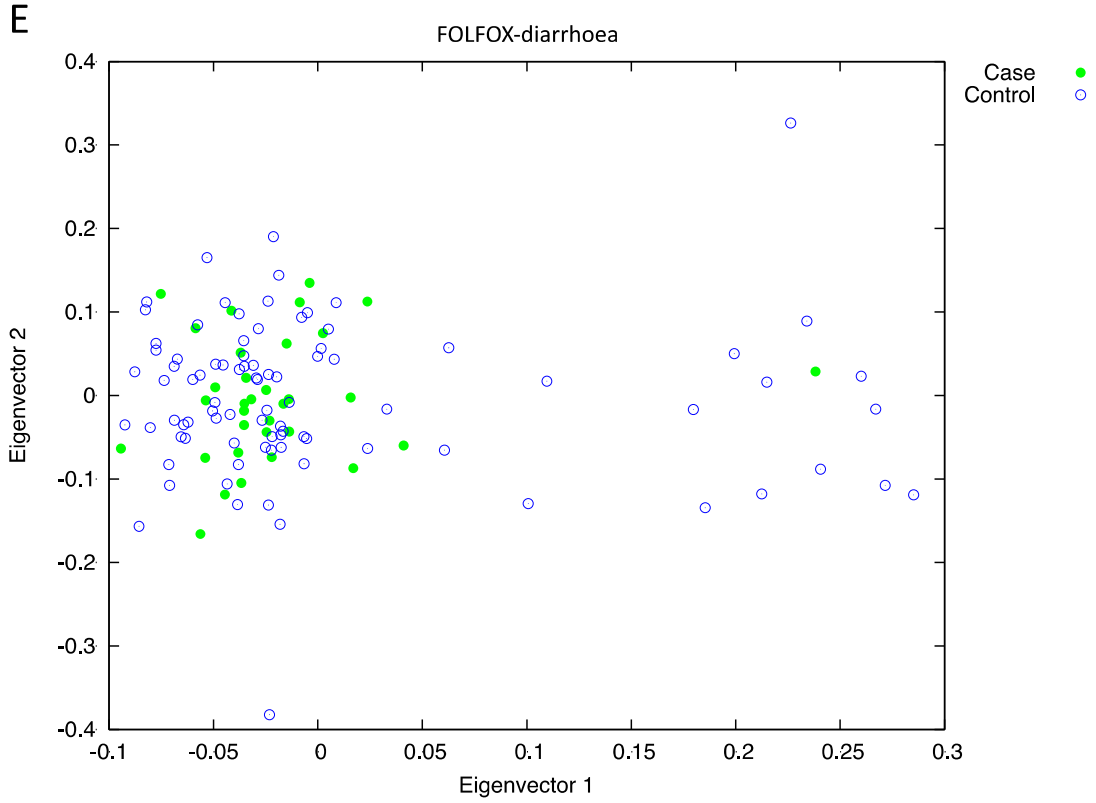
C


5FU-mucositis
Case
Control
Eigenvector 2
Eigenvector 1
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2

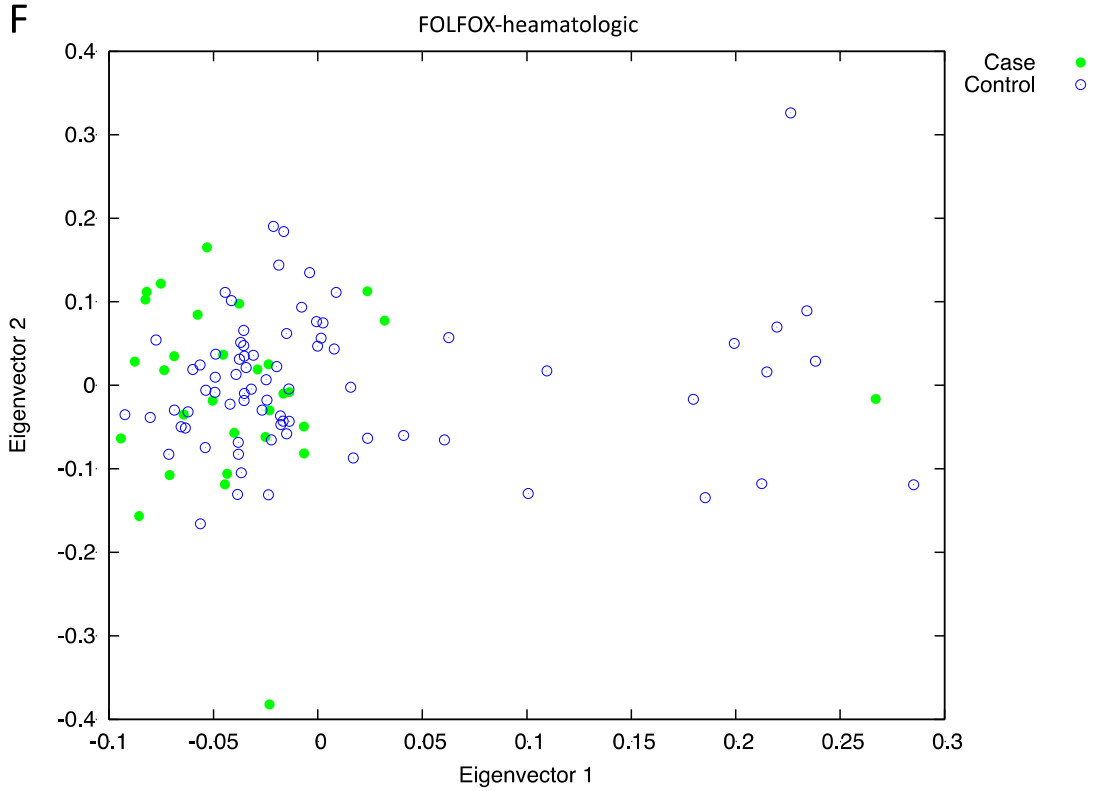
D


5FU-nausea
Case
Control
Eigenvector 2
Eigenvector 1
0.5
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.5
-0.5
-0.4
-0.3
-0.2
-0.1
0
0.1
0.2

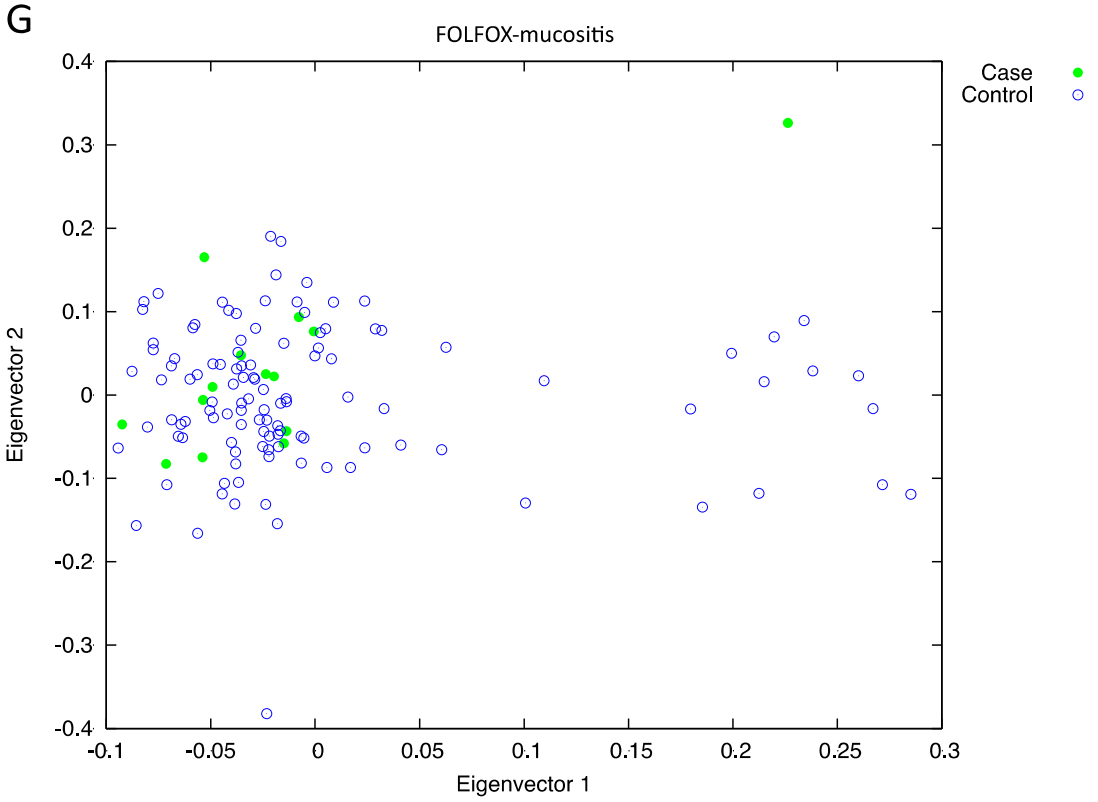
E


FOLFOX-diarrhoea
Case
Control
Eigenvector 2
Eigenvector 1
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3

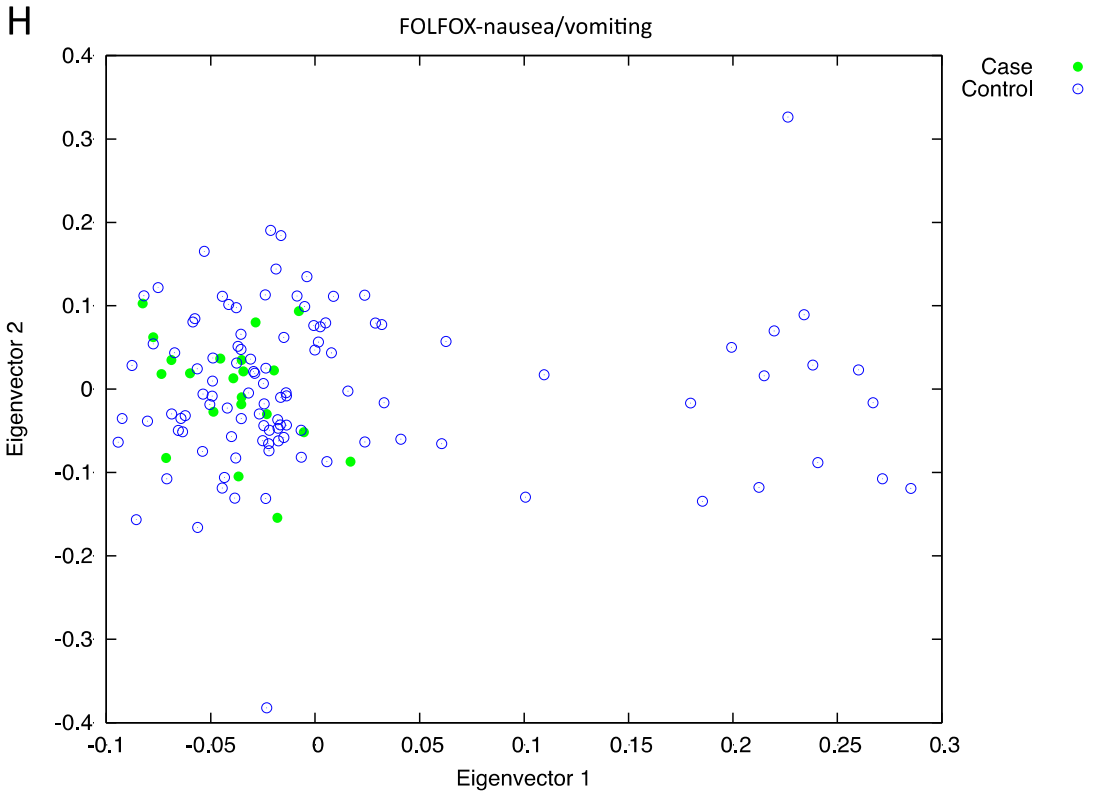
F

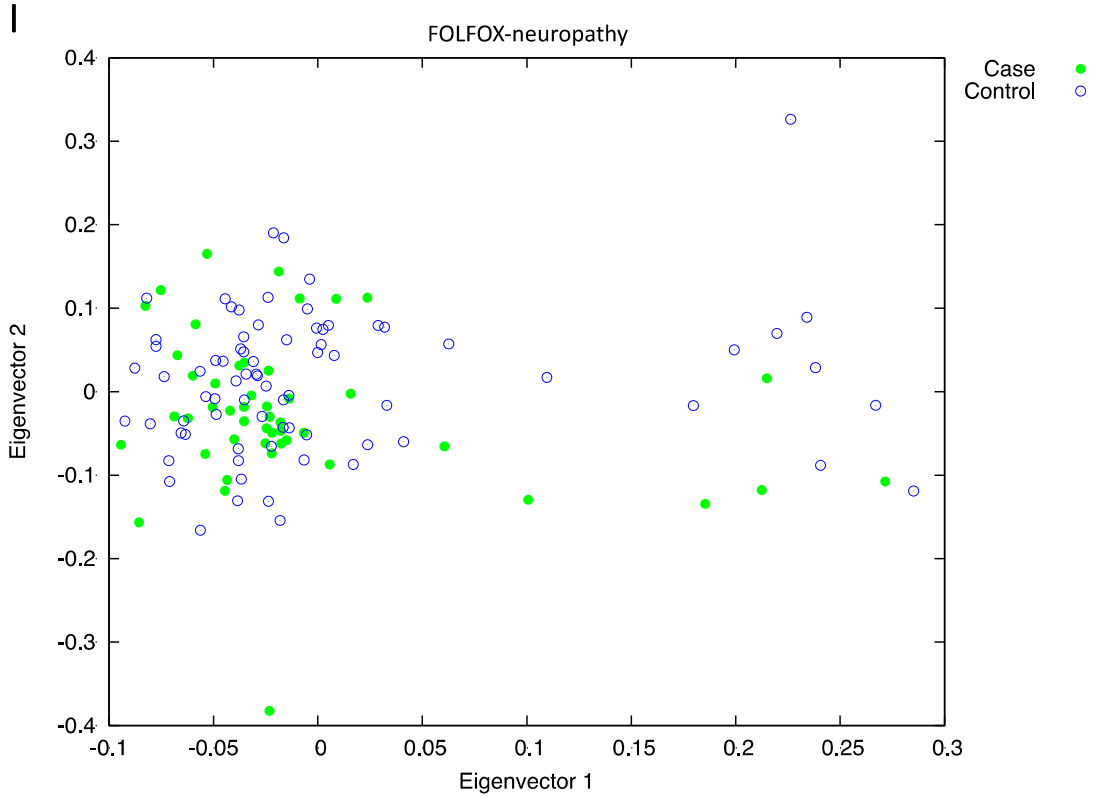

FOLFOX-heamatologic
Case
Control
Eigenvector 2
Eigenvector 1
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3

G

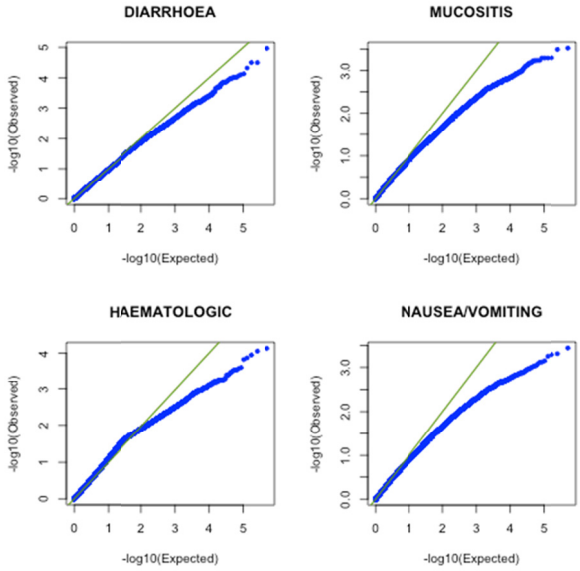

FOLFOX-mucositis
Case
Control
Eigenvector 2
Eigenvector 1
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3

H

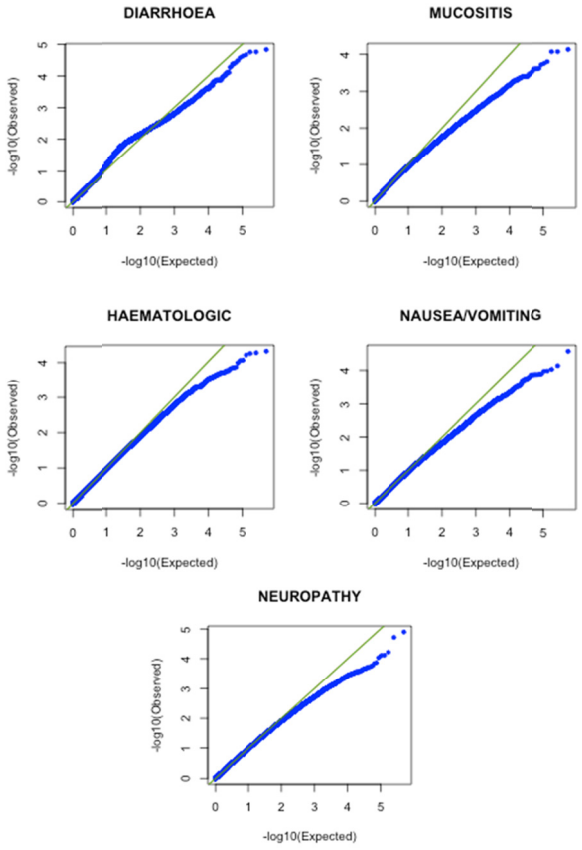

FOLFOX-nausea/vomiting
Case
Control
Eigenvector 2
Eigenvector 1
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3


FOLFOX-neuropathy
Case
Control
Eigenvector 2
Eigenvector 1
0.4
0.3
0.2
0.1
0
-0.1
-0.2
-0.3
-0.4
-0.1
-0.05
0
0.05
0.1
0.15
0.2
0.25
0.3

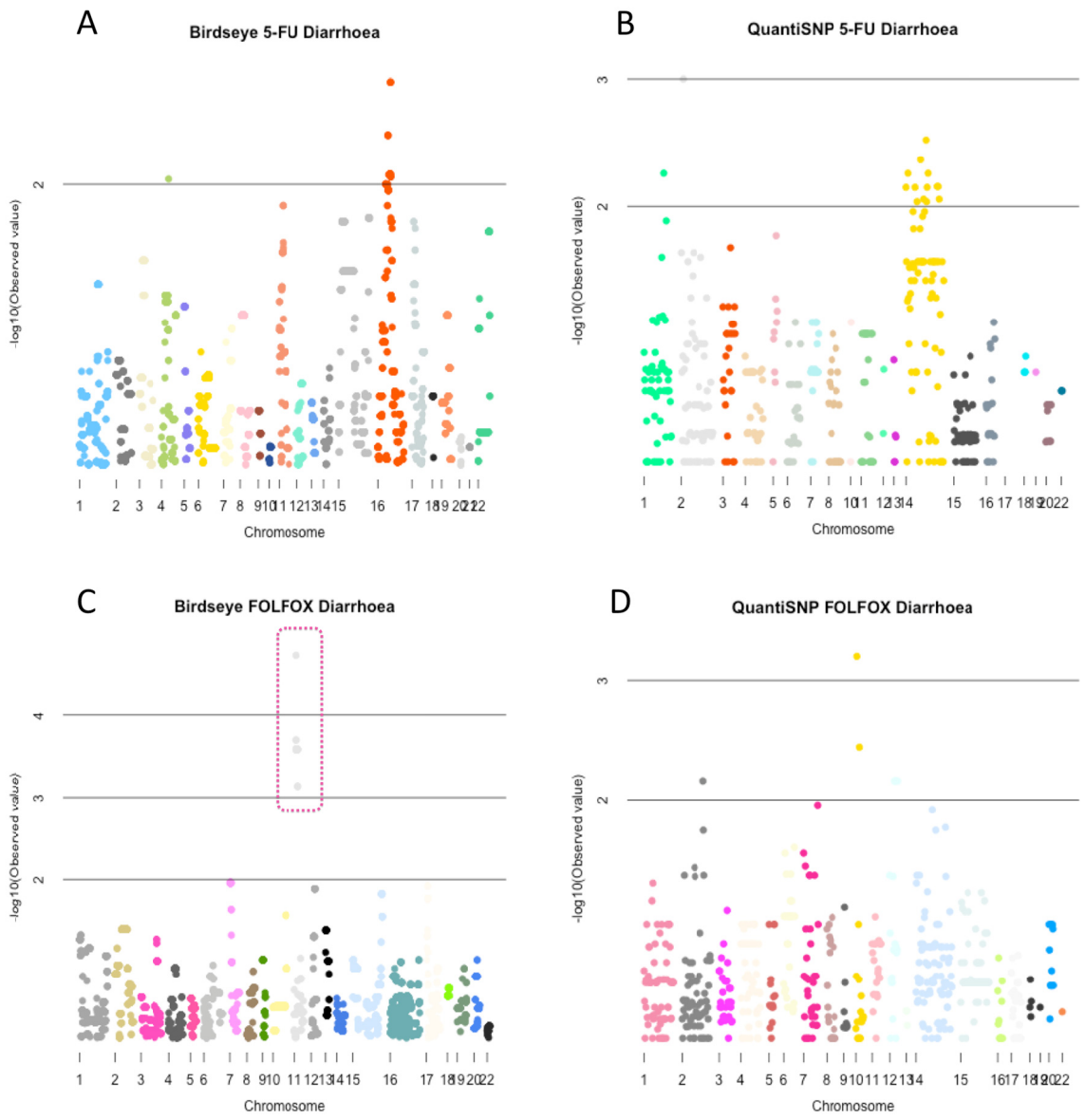
Supplementary Figure 2. Q-Q plots for 5-FU. Distribution of association p-values for 5-FU and each of the ADR phenotypes.



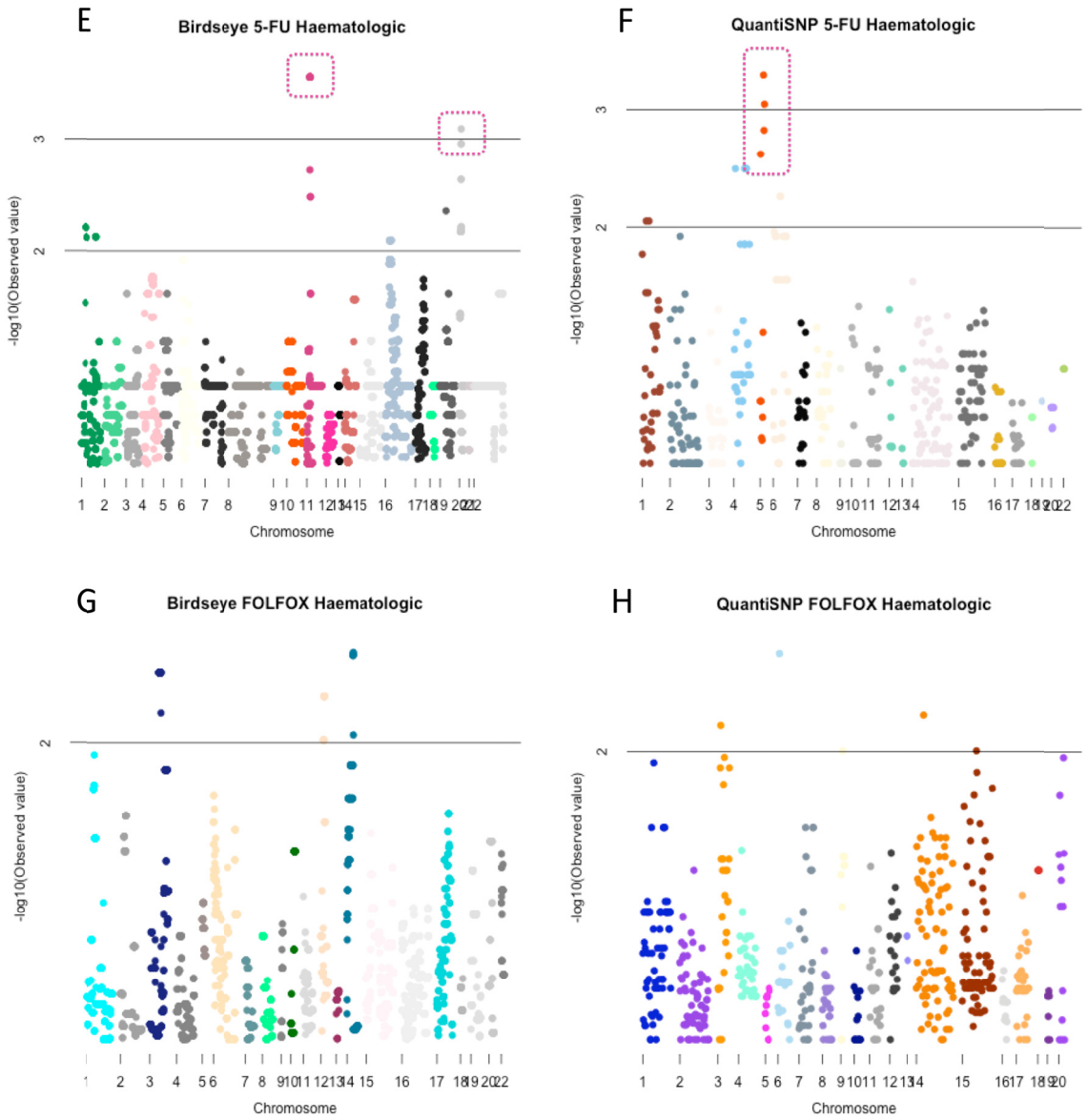
Supplementary Figure 3. Q-Q plots for FOLFOX. Distribution of association p-values for FOLFOX and each of the ADR phenotypes.



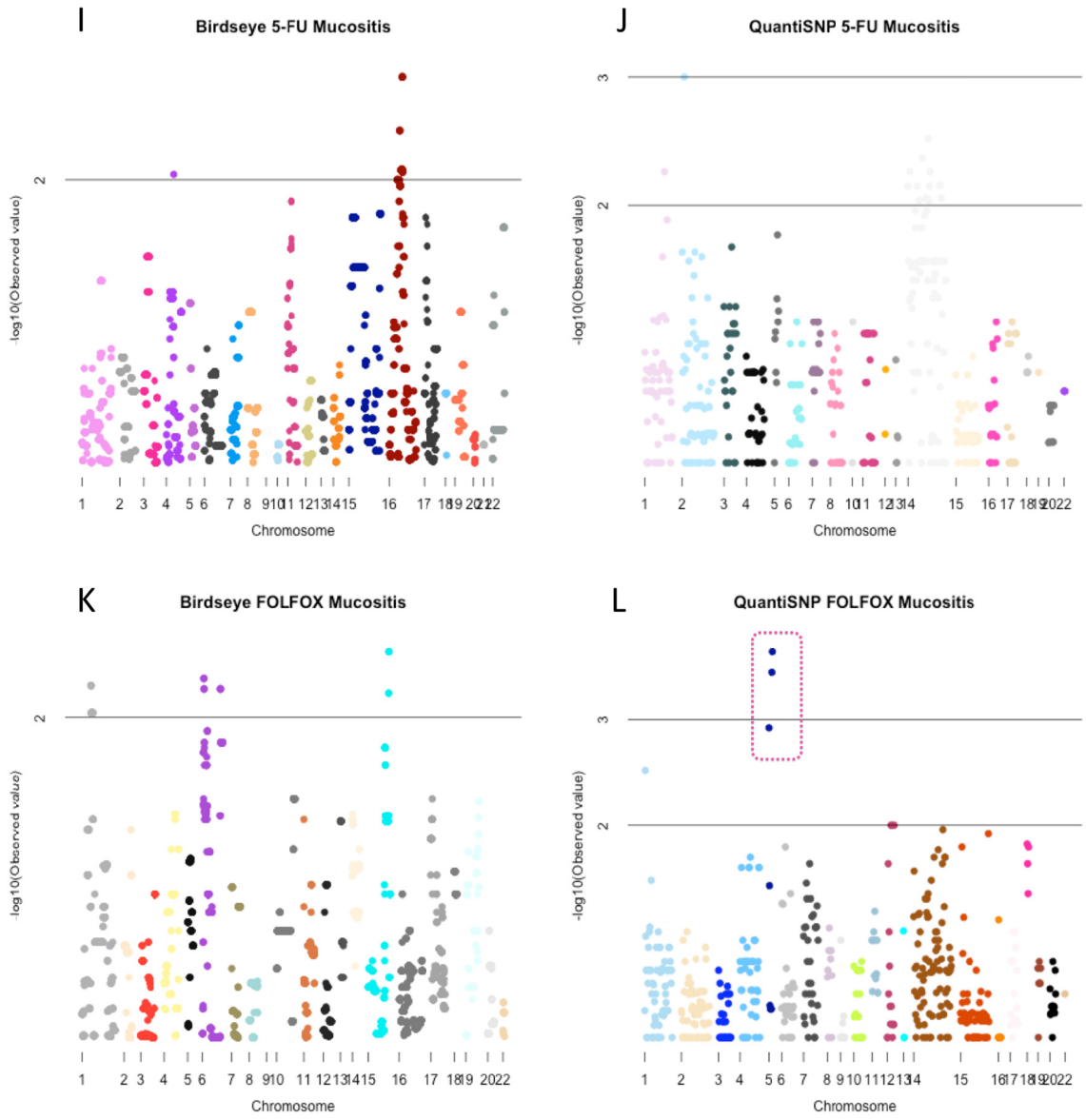
Supplementary Figure 4. Manhattan plots of association for the CNV analyses. Both Birdseye and QuantiSNP results are shown for 5-FU and FOLFOX. The CNVs replicated on stage 2 are highlighted in dashed pink. Top: 5-FU; bottom: FOLFOX; left: Birdseye; right: QuantiSNP. A-D: Diarrhoea.



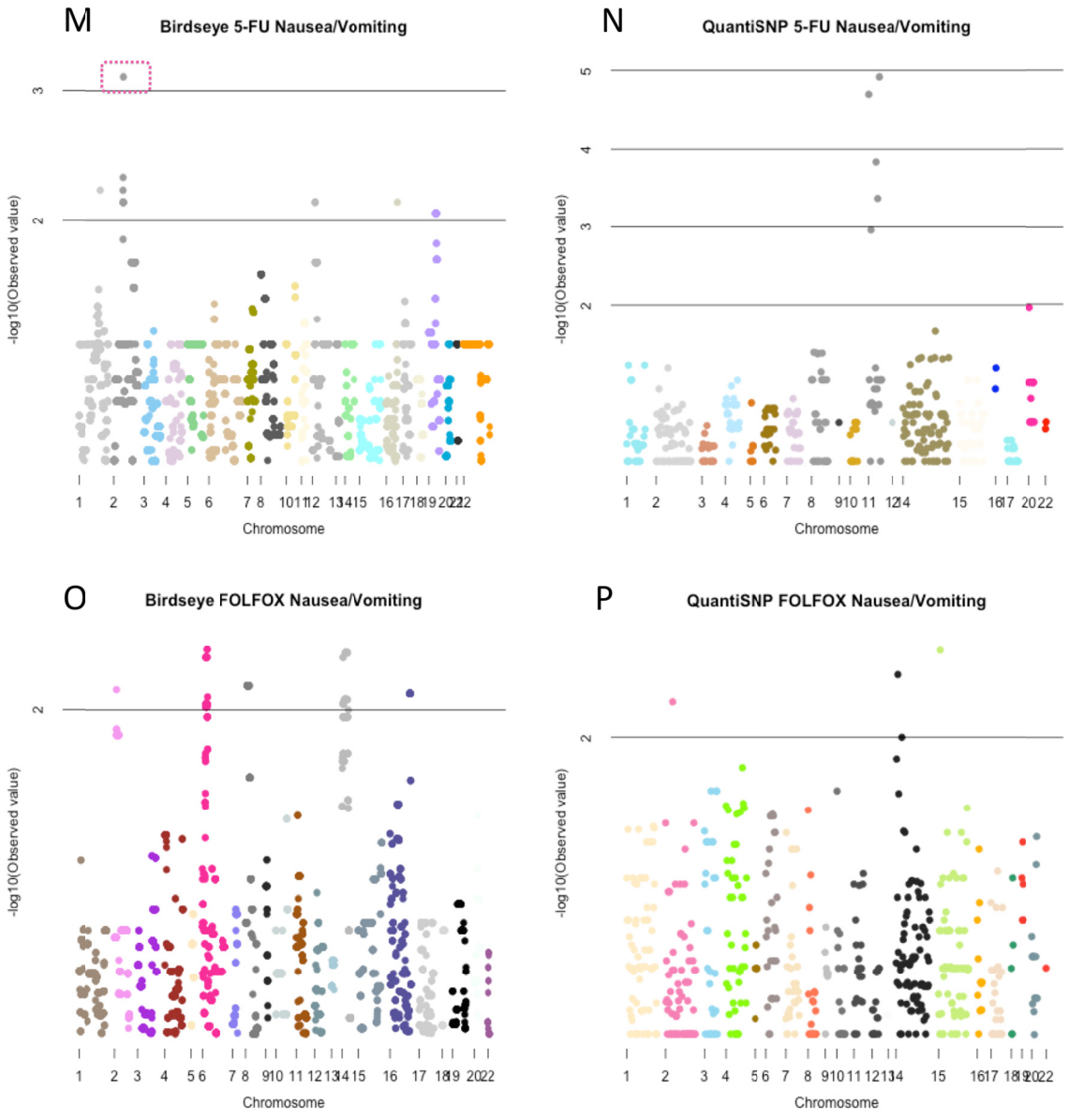
Supplementary Figure 4. (Continuation I). Top: 5-FU; bottom: FOLFOX; left: Birdseye; right: QuantiSNP. E-H: Haematologic.



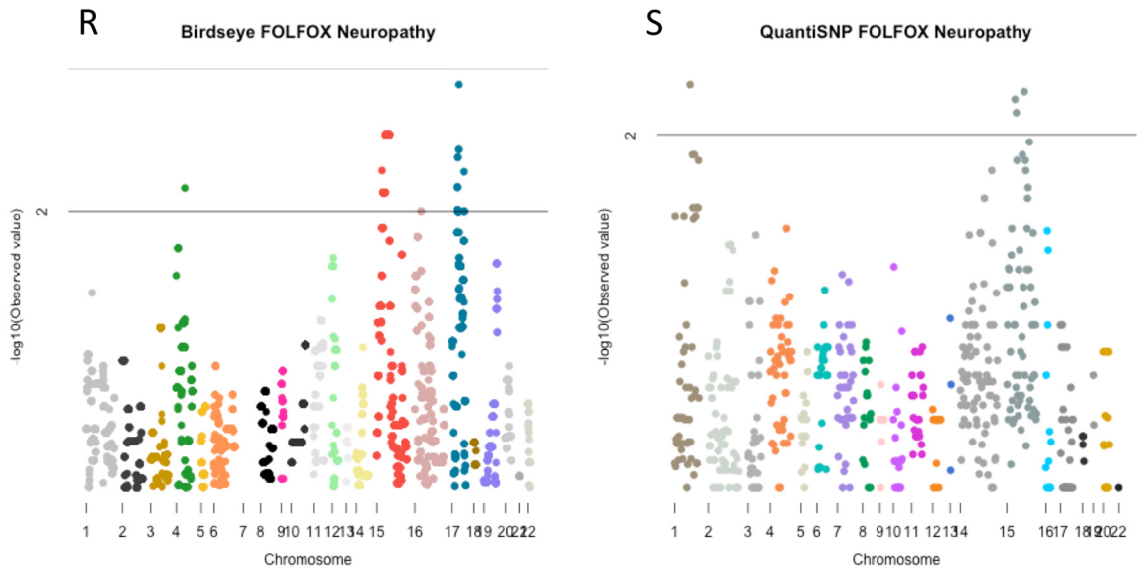
Supplementary Figure 4. (Continuation II). Top: 5-FU; bottom: FOLFOX; left: Birdseye; right: QuantiSNP. I-L: Mucositis.



Supplementary Figure 4. (Continuation III). Top: 5-FU; bottom: FOLFOX; left: Birdseye; right: QuantiSNP. M-P: Nausea/Vomiting.



Supplementary Figure 4. (Continuation IV). FOLFOX; left: Birdseye; right: QuantiSNP. R-S: Neuropathy.

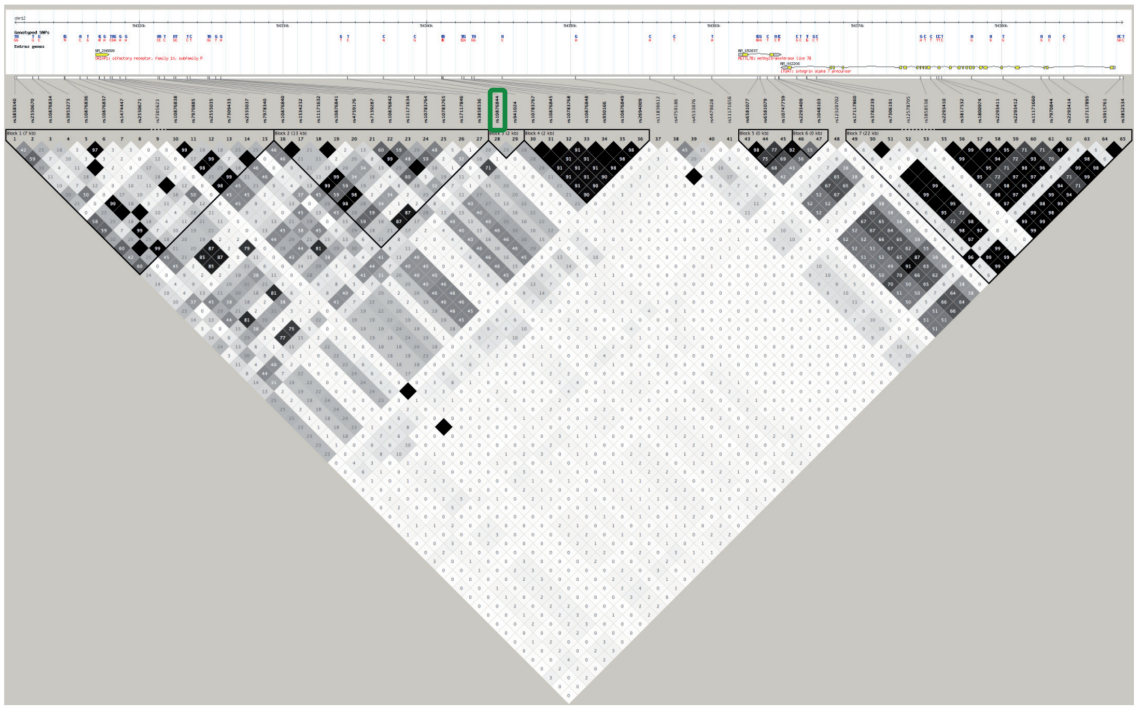


Supplementary Table 1. Association values for each stage and the pooled analysis for the top 5/10 loci selected for replication on stage II. Both SNP and CNV markers are depicted.

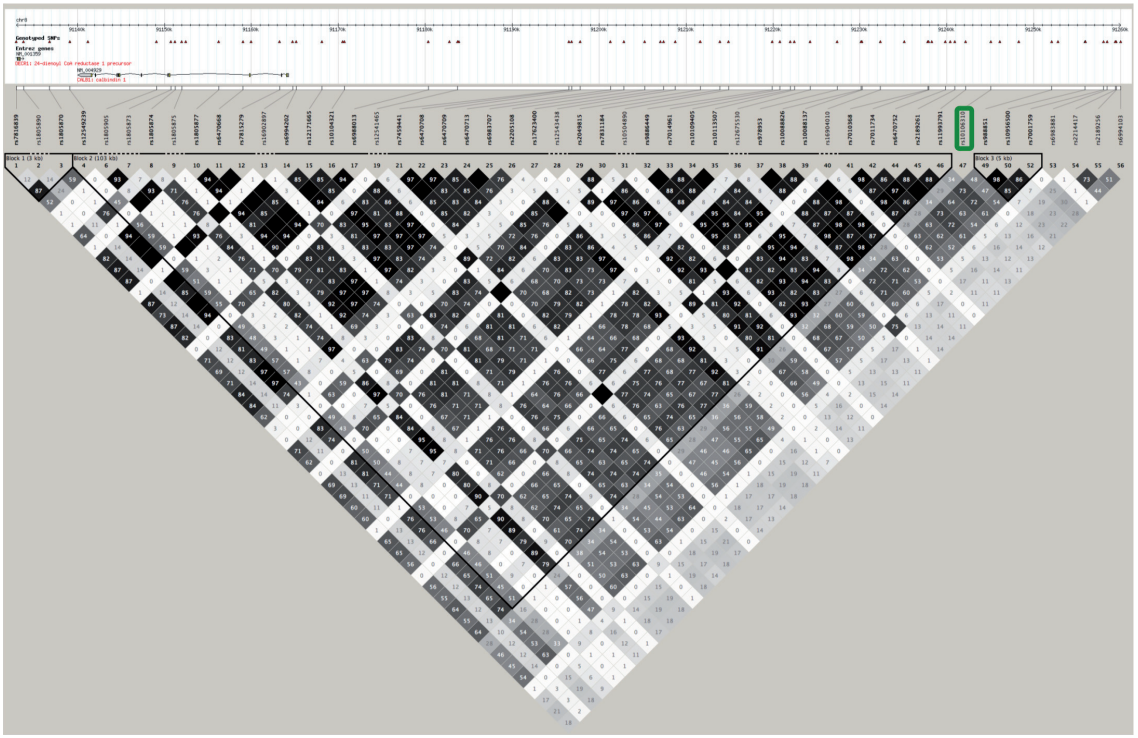
CHR	SNP	POSITION	MINOR ALLELE	P-VALUE PHASE I	P-VALUE PHASE II	P-VALUE POOLED	PHENOTYPE (ADR)
2	rs6713755	223746679	G	3.170E-05	0.028	0.052	5-FU-diarrhoea
10	rs887339	29545130	G	1.076E-05	0.638	0.614	5-FU-diarrhoea
12	rs10876844	54336973	T	4.784E-05	0.023	0.010	5-FU-diarrhoea
12	rs10784749	67708296	C	1.174E-04	0.621	0.017	5-FU-diarrhoea
20	rs4813881	8916642	A	7.854E-05	0.525	0.374	5-FU-diarrhoea
3	rs9861291	180274372	G	7.473E-05	0.233	0.688	5-FU-haematologic
5	rs10055794	171054516	C	8.932E-05	0.380	0.955	5-FU-haematologic
16	rs7184580	77418447	C	1.138E-04	0.983	0.551	5-FU-haematologic
18	rs1943423	55641304	C	2.507E-04	0.684	0.259	5-FU-haematologic
19	rs4805974	39248137	A	1.358E-04	0.679	0.995	5-FU-haematologic
2	rs2627043	179290782	T	5.973E-04	NA	NA	5-FU-mucositis
3	rs16857540	175383269	C	5.130E-04	0.102	0.020	5-FU-mucositis
8	rs2465403	120160008	G	3.282E-04	0.853	9.426E-03	5-FU-mucositis
13	rs927553	23706704	G	3.049E-04	0.209	0.957	5-FU-mucositis
18	rs670454	8682920	C	6.997E-04	0.605	5.084E-03	5-FU-mucositis
2	rs10182133	75710000	G	5.612E-04	0.463	0.058	5-FU-nausea/vomiting
4	rs2060645	183344417	T	5.268E-04	0.139	0.014	5-FU-nausea/vomiting
4	rs6815391	189161973	A	7.086E-04	0.523	0.079	5-FU-nausea/vomiting
10	rs7094179	14912433	T	3.709E-04	0.781	0.391	5-FU-nausea/vomiting
13	rs9300811	102834312	G	5.014E-04	0.934	0.112	5-FU-nausea/vomiting
2	CNV-rs10179790	34594505	T	6.300E-04	0.351	NA*	5-FU-nausea/vomiting
5	CNV-rs2387715	180293872	A	5.100E-04	NA	NA	5-FU-haematologic
11	CNV-rs4466833	49590126	T	2.800E-04	0.475	0.265	5-FU-haematologic
20	CNV-rs2209313	1547142	T	3.300E-05	0.754	0.754	5-FU-haematologic
2	rs4971347	933636	A	2.701E-05	0.364	0.394	FOLFOX-diarrhoea
2	rs250944	234661957	C	2.182E-05	0.967	0.866	FOLFOX-diarrhoea
11	rs1865282	4787228	C	1.489E-05	0.256	0.333	FOLFOX-diarrhoea
12	rs1347851	89091109	C	4.227E-05	0.257	0.205	FOLFOX-diarrhoea
21	rs2282469	34151289	C	3.820E-05	0.333	0.245	FOLFOX-diarrhoea
2	rs17626122	206182257	G	5.950E-05	0.037	4.851E-04	FOLFOX-haematologic
8	rs10106310	91245050	C	5.702E-05	0.057	1.193E-04	FOLFOX-haematologic
13	rs7325568	39716284	C	6.608E-05	0.660	0.067	FOLFOX-haematologic
14	rs2798389	82120903	G	9.292E-05	0.930	0.183	FOLFOX-haematologic
15	rs4243761	23833884	G	5.127E-05	0.471	0.223	FOLFOX-haematologic
1	rs520227	188835935	C	8.459E-05	0.476	0.386	FOLFOX-mucositis
2	rs4128317	29513358	C	7.346E-05	0.252	5.410E-03	FOLFOX-mucositis
2	rs839533	212516352	T	1.585E-04	0.096	4.426E-03	FOLFOX-mucositis
14	rs17098912	99174082	A	2.480E-04	0.580	0.357	FOLFOX-mucositis
19	rs7255865	21596808	C	1.778E-04	0.543	0.552	FOLFOX-mucositis
1	rs2389972	85963254	G	1.069E-04	0.970	0.171	FOLFOX-nausea/vomiting
1	rs10158985	224117232	A	1.365E-04	0.140	6.863E-03	FOLFOX-nausea/vomiting
6	rs851974	152013380	C	0.00009367	0.741	0.0714	FOLFOX-nausea/vomiting
8	rs2739171	134167847	T	1.043E-04	0.992	0.156	FOLFOX-nausea/vomiting
9	rs724975	120828533	A	1.349E-04	0.304	0.609	FOLFOX-nausea/vomiting
1	rs1555070	175039581	A	2.437E-04	0.976	0.130	FOLFOX-neuropathy
2	rs1097701	129863806	A	1.881E-04	0.623	0.157	FOLFOX-neuropathy
3	rs447978	121973776	C	1.354E-04	0.494	0.348	FOLFOX-neuropathy
5	rs9312960	257082	T	1.910E-04	0.044	0.506	FOLFOX-neuropathy
5	rs4957020	367346	A	1.909E-04	0.875	0.114	FOLFOX-neuropathy
5	rs12188653	101411702	C	1.425E-04	0.468	0.028	FOLFOX-neuropathy
5	rs6863960	115023680	G	2.269E-04	0.971	0.271	FOLFOX-neuropathy
11	rs17718902	17741283	G	1.790E-04	0.090	0.771	FOLFOX-neuropathy
11	rs1944118	110857242	A	1.268E-05	0.351	0.438	FOLFOX-neuropathy
17	rs11080058	23766187	T	2.220E-04	0.322	3.589E-03	FOLFOX-neuropathy
5	CNV-rs2387715	180293872	A	2.300E-04	NA	NA	FOLFOX-mucositis
11	CNV-rs10838648	5772861	G	2.600E-04	0.219	0.219	FOLFOX-diarrhoea

NA: not available for genotyping design reasons; NA: rs10179790 was not included in the Affymetrix 6.0 array, and thus no pooled analysis could be performed.*

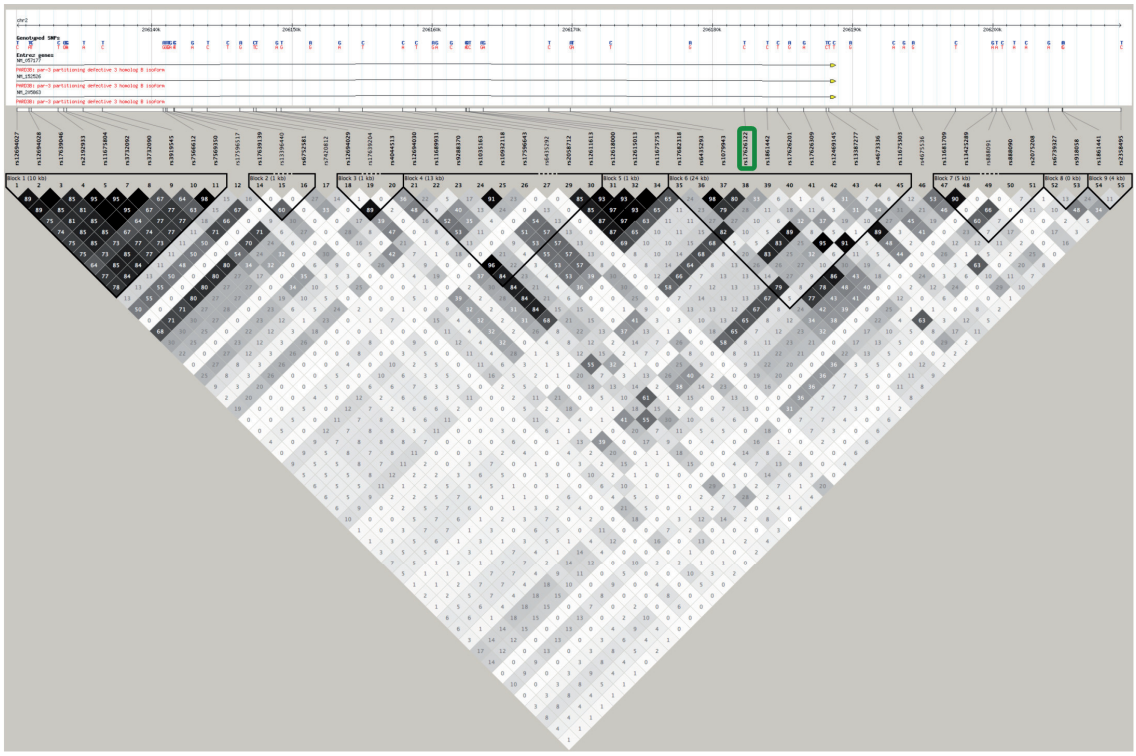
Supplementary Figure 5. LD and rs10876844. LD structure for the block containing rs10876844; the SNP is highlighted in green. R-squared values are shown.



Supplementary Figure 6. LD plot for the region surrounding rs10106310. LD structure for the block containing rs10106310; the SNP is highlighted in green. R-squared values are shown.



Supplementary Figure 7. LD block at rs17626122. LD structure for the block containing rs17626122; the SNP is highlighted in green. R-squared values are shown.



DISCUSSION AND FUTURE PERSPECTIVES

Colorectal cancer is one of the most important forms of neoplasia nowadays, and an important health issue worldwide[1]. Considered a complex disease, the development of CRC is determined by an interplay of both genetic and environmental factors. The genetic component in CRC has been estimated to be around 35% by twin studies[2]. However, only a small proportion of this genetic susceptibility has been identified. Rare, high-penetrance mutations, causing the so-called hereditary CRC syndromes, explain only 5% of the disease cases, whereas moderately penetrant common risk variants are though to account for around 7% of the rest, with the remaining susceptibility variation yet to be discovered.

Thus, one of the aims of this thesis has been the identification of other, yet unidentified, moderate/low-penetrance alleles that may account for at least part of the remaining genetic risk attributed to CRC. In this framework, case-control association studies (CCAS) have been the most widely used implementation to search for such variants. CCAS may be approached in two ways: candidate-gene analyses, screening polymorphisms in genes that are related to carcinogenetic events, and Genome-Wide Association Studies (GWAS), offering a hypothesis free approach to explore the overall variability of the genome[3].

Candidate-gene studies.

The study of model animals, and particularly mice, has been very important in the unravelling of the physiological and genetic processes that occur during CRC tumour development[4]. As a matter of fact, studies in rodent were able to satisfyingly identify 15 QTL loci conferring susceptibility to developing CRC (*Scc* or Susceptibility to Colorectal Cancer)[5-7]. Hence, we decided to investigate the human-mouse syntenic regions of the 15 Scc QTLs, searching for relevant genes that could be functionally

related to carcinogenetic processes. *PTPRJ*, the gene responsible for the Scc1 QTL signal, had already been linked to CRC in humans, and an association study had already been carried out with ambiguous results[8].

Selection of candidate genes was approached by inspection of enriched expression in primary affected tissues in humans[9]. Once the gene list was obtained, we screened a number of SNP markers for evidences of any new potential association signals. By these means, we were able to identify a region defined by rs954353, located in the 5´UTR region of the *CYR61* gene. *In silico* analysis of this locus identified the presence of transcription factor binding sites (TFBSs) at either sides of rs954353. Even when direct sequencing of these two did not succeed in detecting any variants within the TFBSs themselves, we did discover a 6bp expansion polymorphism (3-4 repeats) that was highly correlated with rs954353. Despite confirmation of this association signal being ineffective in a second-stage follow-up, we believe the reduced sample size from our study may have been a key factor in the unproductive replication, and therefore larger cohorts may be needed to fully ascertain the relationship between this polymorphism and CRC. Unluckily, neither rs954353, nor the initial rs12086058 SNP were well-tagged in the Affymetrix 6.0 array (best proxy SNP rs12072027; r^2=0.288); thus, we could not obtain any new information regarding the potential implication of these variants with CRC susceptibility.

Furthermore, we also examined part of the variability in the genes belonging to two important signalling pathways in CRC development: Wnt and BMP. Both of these have been proven essential in the colorectal neoplasic sequence: Wnt is one of the principal CRC tumorigenesis pathways, and contains genes that have been long known to cause CRC syndromes (*APC*, for instance)[10], whereas BMP genes have proto-oncogenic roles in the pathogenesis of CRC and other cancers[11]. Although we were unable to identify

any new risk variants within our study, the importance of some of these genes in CRC susceptiliby, is undeniable. Three different genes in either pathway (*BMP4*, *BMP2* and *GREM1*) have been specifically described in GWA studies as relevant in CRC susceptibility[12] (Carvajal-Carmona *et al.*, PloS Genet *in press*). We consider our inability to detect these effects a limitation due to sample size restrictions. Additionally, our variant selection criteria included only missense and regulatory changes within genic regions, and thus they do not offer a whatsoever comprehensively coverage of the selected genic regions.

GWAS

Although candidate-gene approaches have been extensively performed, the conclusions they have provided on the genetic nature of complex diseases have been scarce. Unsuitable sample sizes, liberal p-values for association calling, lack of appropriate quality control measures (such as multiple-testing corrections) and the restrictions of a functional hypothesis have mostly resulted in a notable impossibility to replicate positive findings[13]. A change of direction was eminently experimented when CCAS switched into the performance of GWAS. This assumption-free approach has indeed been proven to be very advantageous, and it has constituted a meaningful milestone in the unravelling of the genetic basis of common diseases, with 14 new susceptibility loci identified for CRC[12,14].

In this context, we carried out a GWAS analysis on 881 cases and 667 controls in a Spanish cohort. Our study favourably identified a new susceptibility variant at 8p12 (rs11987193). The importance of this finding may be coupled with the fact that this also is the first work to describe a CRC GWAS analysis in a Southern-European population.

During the association analysis, we were able to detect 64 variants at 24 genomic loci that were associated with CRC susceptibility in the first stage of the analysis after extensive quality control measures. Nevertheless, replication of these signals could not be directly achieved by search for association signals in the nearby regions of all 24 loci in the British CORGI population. Although at first this lack of replication would seem like the signals could correspond to type I errors of the analysis, we found reasonable evidences that some of the SNPs behaved differently in the EPICOLON and CORGI populations. Consequently, we decided to genotype the 10 divergent SNPs in a second cohort of Spanish samples.

One of the markers, rs11987193 at 8p12, was positively replicated in this second stage. Although p values in the replication stage were not exceedingly good (p=0.039; OR=0.847 (0.724-0.991)), pooled analysis of both phases was consistent with the assumption of the T allele of this SNP having a protective effect on CRC risk ((p=2.061x10^{-5}; OR=0.788 (0.706-0.879)). The SNP lies 128kb downstream the *DUSP4* gene, which encodes for a MAP kinase phosphatase that has been implicated in carcinogenetic processes[15]. Members of this phosphatase family of proteins are responsible for MAPK inactivation, thus preventing cell division and differentiation, and trigging TGFß-induced apoptosis[15], thus making them exceptional candidates to harbour susceptibility variants for CRC and other types of solid neoplasia.

We find that the special qualities of this association signal, particularly the divergence in MAFs between EPICOLON and CORGI, could be a reflection of variation in the local LD patterns of this locus leading to differences in the ability of the tagger to capture the real causative variant. This could be a way to explain the inability to detect the association signal in the CORGI cohort, even when SNP features (both MAF and OR) are similar to those of the loci that have already been identified in these British

samples. Even though the European population is presumed to be quite homogeneous, we cannot however discard the possibility that the effect of this variant may have a higher prevalence or even be specific to one of the populations. Although this would indeed be less likely, since most specific variants with these features have been described in considerably divergent populations, similar events have been described in CRC and other diseases[16,17]. If we consider the possibility that these local differences may indeed be happening at other locations in the genome, the screening of other Southern European populations could be a helpful tool for the discovery of new susceptibility variants for CRC and other complex diseases.

Besides the newly-identified variant at 8p12, we notably replicated the association signals for 6 of the 14 already-described loci (8q24, 10p14, 11q23, 12q13, 15q13 and 18q21). No evidences were found for the other 8 loci whatsoever, but ORs were consistent with the bibliography[12,14]. Although this lack of replication could well be due to sample size restrictions, it could also be the case that, same as with our marker at the 8p12 locus, population frequency differences constitute a limitation in our ability to detect these association signals. Surely further analyses in larger cohorts will be decisive in the determination of the specificities of these variants and CRC risk in Southern European populations.

CNVs

Even when SNPs are the most frequent form of genetic polymorphisms, genomic variability exists in very different forms, ranging from single base changes to large chromosomal rearrangements[18]. If this is so, then it is sensible to believe that other kinds of common variation may also be influencing an individual´s susceptibility to

common diseases. In the recent years, the discovery of the presence and abundance of copy-number variants (CNVs) shifted the attention of CCAS towards these forms of genetic markers.

The study of CNVs does not come straightforward however. Copy-number assignment is often a complicated mathematical procedure, and several methodologies have been proposed for analysis, with no agreement on overall standard analytic procedures[19]. Besides, association studies of CNVs suffered a major drawback when some authors proposed that the CNV variability had already been indirectly captured through SNP GWAS[20,21]. This statement was based on the fact that pairwise LD measures for common SNP and CNV markers seemed to be very good, and so, SNPs could be used as taggers in the screening of CNV variation.

Although this relationship may be overall true, there are a couple of factors that should be considered. Firstly, CNV regions are often underrepresented in HapMap because they are experimentally difficult regions to sequence, and thus mostly correspond to the heterochromatic portion that is missing from reference sequences (and thereby are barely covered in commercial arrays). Secondly, when present, SNPs in these regions are often compromised during standard quality control procedures, and so in most of the cases, fail to make it through the association analyses unless they are specifically considered[22]. For these reasons, we believed that an assessment of the CNV common variability was necessary in itself and we performed a GWAS study on CN variation at 1M locations. CNV probes were not directed against specific described regions, but designed to cover the whole of the genome physically. This pattern was optimal for our study, since we wanted to avoid the previous ascertainment biases, and also allowed for new CNV discovery.

To reduce the chances of false positive findings, we implemented the use of two different calling algorithms: Birdsuite´s Birdseye[23] and QuantiSNP v2[24]. By these means, an association signal present in both analyses would have smaller chances of being spurious[25]. Accordingly to the physical design of the array, we found that only a small proportion of the sites were copy-number variable, and an even smaller proportion polymorphic (at a frequency of >5% in the population). However, the overlap between common CNV regions between both algorithms was quite remarkable, thus supporting the accuracy of the calling procedure. Indeed, most of the shared regions had already been described in preceding studies and were included in genomic variation databases[26].

Eleven of the polymorphic CNVs (CNPs) at ten different loci showed considerable evidences of association with the CRC phenotype. These were at 2p22.3, 4p16.1, 6q14.1, 11q11, 15p11.1-q11.1, 15q13.3, 16p11.2-p11.1, 17q12, 17q21.31 and 18q12.2. As has happened many times with SNP GWAS, the analysis of these locations does not directly lead to a biological explanation on how these CNVs affect CRC risk. In fact, most of the genes described within these regions are putative predictions based on ORFs, and even some of the loci are located in gene desserts. The only exception would be 4p16.1, which is linked to the *SLC2A9* glucose transporter. Glycolytic processes have been already linked in the literature with cancer evolution; although this relationship seems to be somatic, the presence of a germline factor that would constitute a growth advantage for transformed tumoural cells could be of relevance in the development of the disease[27].

It is also worth mentioning that for 7 of these loci, the coverage from the SNP part of the array was very low (r^2<0.8). Moreover, for 4 of these locations, there is no appropriate SNP tagger even in HapMap, which enhances the demonstration that not all common CNVs are well-tagged by SNP markers.

Even when the finding of these locations has been consistent both between algorithms and the three subpopulations within EPICOLON II, we must stress that the p-values for most have been very modest, and would not have, for once, resisted multiple testing corrections. Thus, we believe it important that both the presence of these CNPs and copy-number status is double-checked by alternative methodologies, such as MLPA or qPCR, and that association results are replicated in independent sample sets. We also note the fact that, if we expect CNVs to behave similarly to susceptibility SNPs, larger cohorts will be required to detect the subtle effects they may impose on CRC risk.

The stratification problem

Throughout our GWAS studies, both for SNPs and CNVs, we have been faced with a stratification issue. Albeit case and control populations being analogous, the inspection of other variables revealed considerable heterogeneity derived from the hospital of origin of the samples. This effect divided the population into three subgroups: samples from the Meixoeiro hospital (GAL samples; n=366), samples from the Donostia hospital (VAS samples; n=167) and all other samples (REST subpopulation; n=944 samples).

This resulted in the continuous need to adapt analytic procedures to our scenario. For SNP analyses, the problem was mostly overcome by performing a meta-analysis with the results from the three separate populations[28]. This strategy proved successful, since we were able to effectively replicate the association signals at 6 of the 14 described loci. In the case of CNVs, the road was more problematic, with the REST population taking a predominant role and the other two serving as "replicates" of the findings. We found that although associations could not always be replicated in the three groups (probably due to GAL and VAS having such small sample sizes), results were in general

concordant among the three subgroups. Also, the predominant effect of the REST population over the others was very clear, with all association signals driven by results found in this subgroup. Of course, this relationship derives from size effect, since GAL and VAS numbers are considerably smaller than those in REST. Notwithstanding, the subdivision also allows for considerable false positive detection, since only findings that were consistent in the GAL and VAS groups as well were considered as true positives.

Despite the fact that these hospitals were the only collection centres from Galicia (NW Spain) and the Basque Country (N Spain), we do not believe that this effect was due to real genetic differences within the Spanish population. Indeed, there were samples from these hospitals that were not of Galician or Basque origin (as stated by geographical origin of 4 grandparents of the proband) and yet clustered away from the main cloud in the PCA analysis, and some other individuals of Galician or Basque origin that had been collected in other hospitals and still belonged to the main cloud. Thus, this effect clearly corresponds to bias within the collection centre and not to genetic heterogeneity within the Spanish population.

We believe these differences may be the result of differential outcome during the DNA extraction procedure leading to decreased sample qualities or a bias within the sample collection procedure in itself that are reflected on the intensity signals in the array. This would explain the variability observed at the selected quality thresholds for some of the CNPs in our GWAS study, and thereby the lack of concordance between the subgroups in association measures.

Our experience has proven that even when case and control populations seem matching, there are other sources of bias that could lead to increased false positive rates in association findings. The assessment of proper cohort homogeneity is thus of noteworthy importance in the performance of association studies.

Pharmacogenomics

It is a fact that response to drug administration is highly variable. Given any medical treatment, there is a considerable percentage of individuals that is expected not to respond, and a similar fraction that will develop adverse drug reactions (ADRs)[29]. The distribution of this variation has been seen to approach normality, which is highly suggestive of the underlying genetic causes following a polygenic model analogous to the one applied to complex diseases[30]. Considering this, it would be thus most certainly appropriate to believe that common genetic variants may play a relevant role in the determination of the outcome. Knowledge on such factors driving response would enable individualisation of the treatments by optimisation of drug dosages and minimisation of toxic effects.

The identification of these genetic markers has been the purpose of pharmacogenetic studies for many years now. Up to very recently, these have relied on candidate-gene approaches investigating the relationship between genetic polymorphisms in drug detoxification enzymes or membrane transporters with toxicity responses for several pharmaceuticals. Nevertheless, the development in high-throughput genotyping technologies and the recent accomplishements in GWAS studies in complex diseases have encouraged pharmacogenetic studies to step aside these hypothesis-based strategies and go genomic[31].

In this context, we performed a GWAS on 221 CRC patients that had received chemotherapy treatment with two of the most commonly administered agents: 5-fluorouracil (5-FU) and the combined forms of oxaliplatin (FOLFOX). By investigating both SNP and CNV variation, we have been able to identify 11 variants (rs16857540, rs2465403, rs10876844, rs10784749, rs670454, rs10158985, rs4128317, rs17626122, rs839533, rs10106310 and rs11080058) that show considerable evidences of association

with a series of ADRs in a two-stage association study. Of these all, again some appear to be located nearby promising genes (such as rs10876844 and *METTL7B*, rs10106310 and *CALB1* and rs17626122 and *PARD3B*), whereas for others their biological implications remain unclear. Granting that significant findings in GWAS do often not directly provide for a biological on the changes resulting in response variability, they do provide for quantitative measures on how the variants affect outcome results. Thus, the description of markers modifying an individual´s risk to developing ADRs or its ability to respond to a certain treatment could be an essential factor to take into account before treatment decision.

It had been proposed by some authors that, unlike heritability to common diseases, variants underlying pharmacogenetic traits would most probably have large ORs[32]. This assumption was based on the belief that, since drug administration is a recent proceeding on human populations, selection would not have had time to act upon such variants. Our study proves however, that at least for the genetics of 5-FU and FOLFOX toxic reactions, this does not appear to be so. Although detected ORs for these 11 variants are slightly higher than most of the described for disease susceptibility, risk distribution in our study clearly supports a more moderate effect of the genetic markers on toxic outcome. Moreover, the fact that drug administration is contemporary does not necessarily exclude the possibility that selection has been acting on those genomic locations for some other reasons.

Considering this, we would expect that the features of the yet unidentified associated variants resemble those that have been described for disease susceptibility. In this context, we are aware that sample size is a considerable limitation of our own study. The number of samples used in stage I is clearly not large enough to detect even moderately penetrant variants, as is clearly reflected in the quantile-quantile plots.

Hence, it is likely that even when we have been able to successfully identify 11 variants, many may have been missed. It is therefore of great importance that further pharmacogenomic studies are performed on appropriately powered cohorts.

The power restriction consequence appears to have been particularly important in CNV analysis, where only very few markers produced significant association values in stage I of the analysis. Additionally, SNP tagging of CNV variants is not always possible, and there were other interesting CNV loci that could regretfully not be assessed in the second stage by classical SNP genotyping. Replication may therefore have to be accomplished by other molecular methods, such as the previously mentioned MLPA and qPCR. Since, pharmacogenetic traits have been described to be susceptible to allele dosage changes[33,34], we believe that some of these unreplicated CNVs may still be good candidates for strong association signals, and thus further efforts will be made to ascertain the relationship between these and 5-FU/FOLFOX-related ADRs.

It is interesting to note that, although not significant after the replication analysis, one of the CNV locations identified as potentially associated in first stage analysis for 5-FU and nausea/vomiting (2p22.3) has also come up as a result in the overall CRC susceptibility analysis. A proper biological explanation for a single variant conferring risk to more than one phenotype would most likely be difficult to support, although there have been other reports of such happenings[35]. Notwithstanding, it would also be sensible to believe that this event happening twice could be the result of some unknown variable being more prevalent in the one of the nausea/vomiting and general groups, thus yielding a false positive signal (which would be in fact supported by the lack of replication in the pharmacogenomic study).

The dark matter: missing heritability

The percentage of overall heritability explained by the already described loci does still only explain a small fraction of the expected risk for which genetic factors are thought to be responsible, both in CRC susceptibility and in pharmacogenetic studies.

It has been proposed that part of this missing heritability could be explained by the imperfect tagging relationship of the markers identified and the real causative variants, which would have resulted in an underestimation of the overall risks behind these association signals[36].

However, it is also probably realistic to believe that other forms of genetic variation, not only common alleles, may be as well playing an important role in CRC susceptibility. The hypothesis that multiple rare variants are also important in common diseases had already been proposed some years ago[37] and there has been increasing evidence lately that this may be so (rare CNV reports on schizophrenia, for instance). In this context, CCAS present little power and alternative study methodologies will have to be used for the identification of such variables. Surely the advances of new sequencing technologies will represent a most helpful tool in this matter, and will hopefully identify new causative variants in the determination of CRC development.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010;127:2893-917.
2. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. N Engl J Med 2000;343:78-85.
3. Zondervan KT, Cardon LR. Designing candidate gene and genome-wide case-control association studies. Nat Protoc 2007;2:2492-501.
4. Demant P. Cancer susceptibility in the mouse: genetics, biology and implications for human cancer. Nat Rev Genet 2003;4:721-34.
5. Moen CJ, Snoek M, Hart AA, Demant P. Scc-1, a novel colon cancer susceptibility gene in the mouse: linkage to CD44 (Ly-24, Pgp-1) on chromosome 2. Oncogene 1992;7:563-6.

6. Meunier C, Cai J, Fortin A, et al. Characterization of a major colon cancer susceptibility locus (Ccs3) on mouse chromosome 3. Oncogene 2009;.
7. van Wezel T, Ruivenkamp CAL, Stassen APM, Moen CJA, Demant P. Four new colon cancer susceptibility loci, Scc6 to Scc9 in the mouse. Cancer Res 1999;59:4216.
8. Toland AE, Rozek LS, Presswala S, Rennert G, Gruber SB. PTPRJ Haplotypes and Colorectal Cancer Risk. Cancer Epidemiol Biomarkers Prev 2008;17:2782-5.
9. Brown AC, Kai K, May ME, Brown DC, Roopenian DC. ExQuest, a novel method for displaying quantitative gene expression from ESTs. Genomics 2004;83:528-39.
10. Half E, Bercovich D, Rozen P. Familial adenomatous polyposis. Orphanet J Rare Dis 2009;4:22.
11. Deng H, Makizumi R, Ravikumar TS, Dong H, Yang W, Yang WL. Bone morphogenetic protein-4 is overexpressed in colonic adenocarcinomas and promotes migration and invasion of HCT116 cells. Exp Cell Res 2007;313:1033-44.
12. Houlston RS, Webb E, Broderick P, et al. Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat Genet 2008;40:1426-35.
13. Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. Genet Med 2002;4:45-61.
14. Houlston RS, Cheadle J, Dobbins SE, 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;42:973-7.
15. Keyse SM. Dual-specificity MAP kinase phosphatases (MKPs) and cancer. Cancer Metastasis Rev 2008 ;27(2):253-61.
16. Pittman AM, Broderick P, Sullivan K et al. CASP8 variants D302H and _652 6N ins/del do not influence the risk of colorectal cancer in the United Kingdom population. Br J Cancer. 2008;98(8):1434-6.
17. Tosa M, Negoro K, Kinouchi Y et al. Lack of association between *IBD5* and Crohn's disease in Japanese patients demonstrates population-specific differences in inflammatory bowel disease. Scand J Gastroenterol 2006;41(1):48-53.
18. Conrad DF, Pinto D, Redon R, et al. Origins and functional impact of copy number variation in the human genome. Nature 2009; 464(7289):704-12.
19. Carter NP. Nat Genet 2007;39:S16-21.
20. Hinds DA, Kloek AP, Jen M, Chen X, Frazer KA. Common deletions and SNPs are in linkage disequilibrium in the human genome. Nat Genet 2006;38:82-5.
21. Wellcome Trust Case Control Consortium. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. Nature 2010;464:713-20.
22. Redon R, Ishikawa S, Fitch KR, et al. Global variation in copy number in the human genome. Nature 2006;444:444-54.
23. Korn JM, Kuruvilla FG, McCarroll SA, et al. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. Nat Genet 2008;40:1253-60.

24. Colella S, Yau C, Taylor JM, et al. QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. Nucleic Acids Res 2007;35:2013-25.
25. Pinto D, Marshall C, Feuk L, Scherer SW. Copy-number variation in control population cohorts. Hum Mol Genet 2007;16 Spec No. 2:R168-73.
26. Zhang J, Feuk L, Duggan G, Khaja R, Scherer S. Development of bioinformatics resources for display and analysis of copy number and other structural variants in the human genome. Cytogenetic and genome research 2006;115:205-14.
27. Gatenby RA, Gillies RJ. Why do cancers have high aerobic glycolysis? Nature Reviews Cancer 2004;4:891-9.
28. Petitti DB. Meta-analysis, decision analysis, and cost-effectiveness analysis: methods for quantitative synthesis in medicine. Oxford University Press, USA, 2000.
29. Vesell ES. Genetic and environmental factors causing variation in drug response. Mutat Res 1991;247:241-57.
30. Crowley JJ, Sullivan PF, McLeod HL. Pharmacogenomic genome-wide association studies: lessons learned thus far. Pharmacogenomics 2009;10:161-3.
31. Goldstein DB, Tate SK, Sisodiya SM. Pharmacogenetics goes genomic. Nature Reviews Genetics 2003;4:937-47.
32. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. Nature 2009;461:747-53.
33. Schimke RT, Kaufman RJ, Alt FW, Kellems RF. Gene amplification and drug resistance in cultured murine cells. Science 1978;202:1051.
34. Moroni M, Veronese S, Benvenuti S, et al. Gene copy number for epidermal growth factor receptor (EGFR) and clinical response to antiEGFR treatment in colorectal cancer: a cohort study. The lancet oncology 2005;6:279-86.
35. Schumacher FR, Feigelson HS, Cox DG, et al. A common 8q24 variant in prostate and breast cancer from a large nested case-control study. Cancer Res 2007;67:2951.
36. Spencer C, Hechter E, Vukcevic D, Donnelly P. Quantifyping the underestimation of relative risks from Genome-Wide Association Studies. PLoS Genetics 2011;7(3):e1001337.
37. Weiss KM, Terwilliger JD. How many diseases does it take to map a gene with SNPs? Nat Genet 2000;26:151-8.

CONCLUSIONS

The conclusions that may be extracted from this work are as follows:

1. Although we could not successfully find enough evidences of new CRC risk variants through candidate-gene strategies, SNP rs954353, upstream the *CYR61* gene, stands out as a potential candidate that, in our consideration, should not be directly discarded. We believe that sample size restrictions may have been considerably limiting in our study and thus further assessment of the relationship between this variant and CRC susceptibility may be needed.

2. Our GWAS study has successfully identified a new low-penetrance CRC variant, rs11987193, at 8p12. The T allele of this SNP shows a protective effect against CRC risk. This locus harbours the *DUSP4* gene, which had been previously related to carcinogenetic processes. Given the MAF discrepancies between the EPICOLON and CORGI cohorts about the association signal, it is likely that there are differences in the ability to tag the causative variant at this locus between Northern and Southern European populations.

3. We have favourably identified eleven CNPs that present substantial evidences of association with CRC. Despite this association signals being highly reliable, further concordant estimations of copy-number status and replication in larger, independent cohorts will be necessary to fully ascertain their relationship with CRC risk.

4. Even when previous studies have described that common copy-number variation is well tagged by SNPs, our findings suggest that this assumption may have been biased by the incomplete coverage of previous array designs. Thus, we encourage CNV

association studies to be performed in order to completely determine the extent to which this type of variants may be responsible for CRC phenotypes.

5. Our GWAS study on the genetic susceptibility to several ADRs and treatment with 5-fluorouracil and the combined forms of oxaliplatin (FOLFOX) has conveniently described 11 new variants that harbour evidences of association with toxicity phenotypes. We consider that GWAS studies constitute a good implementation for the detection of pharmacogenetic variants, and thus praise future works in this direction be made in larger sample sets.

SUMMARY

Heritability in colorectal cancer (CRC) predisposition has been estimated to be around 35% by twin studies. Although ~ 5% of this proportion may be explained by high-penetrance mutations and an additional 7% is thought to be due to the presence of a combination of some of the already-described 16 susceptibility SNPs, there is still a significant fraction of CRC susceptibility that remains unexplained.

On the other hand, there is also considerable variation in the way CRC patients respond to chemotherapy. Besides, the fact that most drugs used in CRC treatment have narrow therapeutic ranges results in the frequent development of adverse drug reactions (ADRs). Hence, the identification of the genetic variation modulating this outcome would be most helpful in both the individualisation of the treatment and the reduction of health costs.

In this context, we have intended to search for new variants that could explain at least a part of the missing heritability in CRC. For this purpose, we have chosen to investigate the most common sources of variability in the genome: SNPs and CNVs.

In the SNP part of the study, we have followed two different approaches: a *candidate-gene* strategy evaluating the polymorphic variation in genes with a potential functional implication in CRC carcinogenesis and a *genome-wide association study*. For the former, we have assayed in separate studies the genes present in the human syntenic regions of the 15 *Scc* (susceptibility to colorectal cancer) mouse loci, and those belonging to two pathways that have been consistently linked to CRC development: Wnt and BMP. For the latter, we have carried out a GWAS in a Spanish cohort. The advantage of this strategy against the candidate-gene one is that there is no *a priori* hypothesis on where the susceptibility loci may be located.

Regarding the CNV study, we have also performed a GWAS scan of the genomic structural variation and its potential implication in CRC neoplasia, using two different copy-number calling algorithms: Birdsuite´s *Birdseye* and QuantiSNP v2.

In the second part of the study, our purpose has been to analyse the relationship between common genetic variation and the development of ADRs after chemotherapy. For this, we evaluated the correlation between two of the most common administered drugs in CRC treatment: 5-fluorouracil and oxaliplatin (FOLFOX) and the presence of ADRs by screening both SNP and CNV markers at a genome-wide level.

The conclusions that may be extracted from this work are as follows:

1. Although we could not successfully find enough evidences of new CRC risk variants through candidate-gene strategies, rs954353 stands out as a potential candidate that, in our consideration, should not be directly discarded. We believe that sample size restrictions may have been considerably limiting in our study and thus further assessment of the relationship between this variant and CRC susceptibility may be needed.

2. Our GWAS study has successfully identified a new low-penetrance variant at 8p12 conferring risk to CRC. This locus harbours the *DUSP4* gene, which had been previously related to carcinogenetic processes. Given the MAF discrepancies between the EPICOLON and CORGI cohorts about the association signal, it is likely that there are differences in the ability to tag the causative variant at this locus between Northern and Southern European populations.

3. We have favourably identified ten common CNV loci that present substantial evidences of association with CRC. Despite this association signals being highly

reliable, further concordant estimations of copy-number status and replication in larger, independent cohorts will be necessary to fully ascertain their relationship with CRC risk.

4. Even when previous studies have described that common copy-number variation is well tagged by SNPs, our findings suggest that this assumption may have been biased by the incomplete coverage of previous array designs. Thus, we encourage CNV association studies to be performed in order to completely determine the extent to which this type of variants may be responsible for CRC phenotypes.

5. Our GWAS study on the genetic susceptibility to several ADRs and treatment with 5-fluorouracil and the combined forms of oxaliplatin (FOLFOX) has conveniently described 11 new variants that harbour consistent evidences of association with toxicity phenotypes. We consider that GWAS studies constitute a good implementation for the detection of pharmacogenetic variants, and thus praise future works in this direction be made in larger sample sets.

El cáncer colorrectal (CCR) es una de las neoplasias más frecuentes, siendo el tercer tipo de cáncer más común en ambos sexos, después del cáncer de mama en mujeres y el de pulmón en hombres. La tasa global de prevalencia del CCR es del 11.5%, aunque la distribución global no es homogénea, con más del 60% de los casos registrados en países desarrollados. Es una enfermedad especialmente prevalente entre los hombres (ratio 1.4:1), con una edad media de diagnóstico bastante alta (72 años). De hecho, casi el 90% de los casos de CCR se detectan a edades superiores a los 50 años, y un 70% de ellos serán mayores de 56 años.

Se considera que, etiológicamente, el CCR, al igual que otras enfermedades comunes, es una enfermedad compleja. Esto implica que la enfermedad surge como resultado de la interacción entre diversos factores ambientales y genéticos. Estudios en gemelos estimaron en la década de los 90 que la contribución hereditaria al CCR es de aproximadamente un 35%. Esta heredabilidad está representada tanto por mutaciones raras de alta penetrancia, como por un número de variantes comunes en la población que confieren un pequeño efecto sobre el riesgo de desarrollar la enfermedad. Las mutaciones de alta penetrancia, responsables de los síndromes hereditarios de cáncer colorrectal (principalmente la Poliposis Adenomatosa Familiar y el Síndrome de Lynch), sólo son capaces de explicar alrededor de un 5% de los casos de CCR. Es por esto que el estudio de las variantes de baja/moderada penetrancia ha adquirido gran importancia en estos últimos años, ya que se espera que sean las responsables de los casos hasta ahora genéticamente inexplicables de CCR.

La estrategia más optimizada y utilizada para la detección de estas variantes de baja penetrancia son los estudios de asociación. En ellos se compara la frecuencia de una determinada variable entre una población de individuos afectos, o casos, y una de sanos, o controles. Diferencias en la prevalencia de una de estas variantes podrían ser

indicativas de que el cambio esté relacionado con la aparición de la enfermedad. El uso de este tipo de estrategias ha sido impulsado en estos últimos años gracias a la compleción del Proyecto Genoma Humano y al desarrollo de las plataformas de genotipado masivo de estos marcadores. Gracias a esto también, se descubrió que una gran parte del genoma humano era variable y que gran parte de esta variación genética se debía a cambios en una única base: los denominados *Single Nucleotide Polymorphisms* o SNPs.

Inicialmente, los estudios de asociación se centraron en la inspección de SNPs existentes en genes cuya función podría estar relacionada con procesos carcinogénicos. Estos estudios de genes candidatos tenían como ventaja la fácil interpretación biológica de los resultados positivos de asociación y fueron usados extensamente en el estudio del CCR y otras enfermedades complejas. Sin embargo, la mayoría de los hallazgos resultaron infructuosos, ya que los resultados obtenidos no podían ser replicados en estudios posteriores. En este marco, el desarrollo del proyecto HapMap y el descubrimiento de la herencia en bloques del genoma permitieron que los estudios de asociación se pasasen a realizar a escala genómica con los denominados estudios de asociación pangenómicos, o GWAS (del inglés *Genome-Wide Association Studies*). Estos permitían el estudio, libre de hipótesis previas, de una gran cantidad de marcadores repartidos a lo largo de todo el genoma. La comercialización de los *arrays* de GWAS supuso un boom en el mundo de la genómica, posibilitando la identificación de un gran número de loci de susceptibilidad en un gran número de enfermedades. En CCR en particular, los GWAS han identificado 16 nuevas variantes de susceptibilidad en 14 loci distintos: rs6983267 en el locus 8q24, rs4939827 en 18q21.1, rs4779584 en 15q13.3, rs3802842 en 11q23.1, rs16892766 en 8q23.3, rs10795668 en 10p14, rs4444235 en 14q22.2, rs9929218 en 16q22.1, rs10411210 en 19q13, rs961253 en

20p12.3, rs6691170 y rs6687758 en 1q41, rs10936599 en 3q26.2, rs11169552 y rs7136702 en 12q13.13 y rs4925386 en 20q13.33. Como era esperable, todas estas variantes tienen efectos muy modestos sobre el riesgo a desarrollar CCR, con *odds ratios* típicamente por debajo de 1,5. Aún cuando los GWAS han tenido bastante éxito en cuanto al descubrimiento de nuevas variantes de riesgo, los 16 SNPs identificados son sólo capaces de explicar un 7% de la heredabilidad atribuida CCR. Este hecho resalta la importancia de la realización de nuevos estudios con mayores números de muestra o incluso en otras poblaciones, que puedan descubrir nuevas variantes que expliquen la proporción restante de heredabilidad genética.

Por otro lado, se conoce desde hace varias décadas que existe una alta variabilidad interindividual en la respuesta a la administración de fármacos. Estas diferencias se pueden deber a múltiples factores, como la edad, el sexo, el estado de salud del paciente o factores genéticos. El estudio de las variantes genéticas que determinan la distinta respuesta a fármacos es el campo de la farmacogenética. Esta ciencia intenta entender las bases genéticas de la variabilidad observada en el resultado terapéutico con el fin de individualizar los tratamientos para mejorar la respuesta del individuo y minimizar los efectos adversos tóxicos. En el caso del cáncer, la gran mayoría de los medicamentos usados en quimioterapia presentan rangos terapéuticos reducidos. Esto implica que la diferencia entre las dosis efectivas del fármaco y aquellas que causan respuestas tóxicas es muy pequeña. Por ello, la individualización de los protocolos de administración de fármacos para cada paciente es de especial relevancia en este ámbito.

Inicialmente, los trabajos en farmacogenética se centraban en el estudio de variantes raras con efectos mendelianos. A pesar de que se han identificado un gran número de variantes con este tipo de herencia, el patrón normal de distribución de la respuesta

poblacional a fármacos sugiere que al igual que sus homólogos patológicos, el resultado de la administración de medicamentos depende de una herencia poligénica. De este modo, los estudios de asociación pasaron a ser también una metodología importante para la detección de nuevos loci. Al igual que en el caso de los estudios de asociación para la identificación de variantes de riesgo a enfermedades, hasta hace muy poco los estudios de asociación en farmacogenética analizaban mayoritariamente variantes en un número muy reducido de genes candidatos. Los genes codificantes de proteínas transportadoras de moléculas o enzimas metabolizadoras son ejemplos de gran parte de los estudios realizados. En los últimos años, la revolución de los GWAS ha llegado también al campo de la farmacogenética, y ya se han empezado a publicar los primeros resultados de este tipo para algunos fármacos concretos. Se espera que el éxito de esta estrategia sea al menos similar al que han tenido los GWAS en enfermedades.

Objetivos

Con estos antecedentes, el trabajo realizado en esta tesis de doctorado tiene dos objetivos principales:

1. La búsqueda de variantes nuevas de susceptibilidad al CCR.
2. 2. El análisis de la variabilidad genética en relación a las diferentes respuestas tóxicas en pacientes de CCR tratados con quimioterapia.

Para la consecución del primer objetivo, hemos interrogado los dos tipos de polimorfismos genéticos más frecuentes en el genoma, los SNPs y los CNVs. En el primer caso, hemos adoptado además dos estrategias diferentes: una de genes candidatos, evaluando la presencia de SNPs en genes con una potencial implicación funcional en la carcinogénesis colorrectal, y un GWAS. En la parte del estudio de genes candidatos, se investigaron en dos estudios independientes los marcadores presentes en

las regiones sinténicas humanas de los 15 loci de susceptibilidad (Scc) identificados en ratones (Capítulo 1), y también los pertenecientes a dos de las rutas más importantes en el desarrollo de neoplasias colorrectales: la ruta Wnt y la BMP (Capítulo 2). Para la parte del GWAS, el análisis se realizó en una cohorte de muestras españolas (Capítulo 3). La ventaja de esta estrategia frente a la primera es que no existen hipótesis *a priori* sobre la localización de los loci de susceptibilidad.

En cuanto al estudio de CNVs, hemos realizado también un GWAS en la variabilidad genómica estructural con el fin de evaluar su posible implicación en la neoplasia colorrectal mediante el uso de dos algoritmos diferentes de identificación de CNVs: Birdseye y QuantiSNP (Capítulo 4).

La segunda parte de la tesis está dirigida al análisis de la relación entre los polimorfismos genéticos comunes y el desarrollo de respuestas adversas a fármacos tras un tratamiento quimioterapéutico. Para ello, evaluamos la correlación entre dos de los medicamentos más usados en la quimioterapia del CCR: el 5-FU y el FOLFOX, y la presencia de respuestas tóxicas mediante el estudio de ambos SNPs y CNVs a nivel genómico (Capítulo 5).

Resultados y discusión.

El estudio de modelos animales, particularmente de ratones, ha sido de gran importancia para el descubrimiento de los procesos, tanto fisiológicos como genéticos, que ocurren durante la transformación tumoral celular. Por ello, la investigación de las regiones sinténicas en humanos de los 15 QTLs de susceptibilidad al CCR suponía una buena estrategia para la búsqueda de nuevas variantes de susceptibilidad. La selección de los genes candidatos se realizó mediante la inspección de tránscritos que presentaban una

expresión enriquecida en tejidos primarios afectados en humanos. Una vez que se obtuvo esta lista de genes, se seleccionaron una serie de marcadores para buscar potenciales señales de asociación. De este modo, identificamos una región definida por el SNP rs954342, localizado en la región 5´UTR del gen *CYR61*, como candidata a modificar el riesgo de padecer CCR. Los análisis *in silico* de este locus identificaron la presencia de dos lugares de unión a factores de transcripción (TFBS, del inglés *Transcription Factor Binding Sites*), uno a cada lado del SNP. La secuenciación directa de esta región no consiguió identificar ninguna variación dentro de las propias secuencias consenso TFBS. Sin embargo, sí detectamos la presencia de un polimorfismo de expansión de 6 pares de bases, con dos alelos predominantes (3 o 4 repeticiones) y que presentaba un alto grado de correlación con rs954353. A pesar de que la señal de asociación no pudo ser confirmada en una segunda fase, creemos que el reducido número muestral del estudio puede haber sido un factor decisivo en la improductividad de la réplica, y por lo tanto se necesitarán estudios en cohortes mayores para determinar con certeza la relación entre este polimorfismo y el CCR.

Además, examinamos en otro estudio independiente la variabilidad en los genes de las rutas Wnt y BMP. Ambas rutas son esenciales en la secuencia neoplásica colorrectal. Wnt es una de las principales vías tumorogénicas, y contiene genes, como por ejemplo *APC*, que han sido relacionados con la aparición de síndromes hereditarios. Los genes de la ruta BMP son predominantemente proto-oncogenes que se han visto involucrados en la patogénesis de CCR y otros tipos de cánceres. A pesar de que no pudimos identificar ninguna variante nueva de riesgo a través de nuestro estudio, la importancia de algunos de estos genes en la susceptibilidad a desarrollar CCR es innegable. Tres genes pertenecientes a estas rutas (*BMP4*, *BMP2* and *GREM1*) han sido específicamente vinculados por GWAS al CCR. Consideramos que nuestra incapacidad de detectar estos

efectos es debida a la limitación impuesta por el número muestral de nuestro estudio. Adicionalmente, nuestra selección de variantes únicamente incluía variantes *missense* y cambios reguladores en las regiones 5´y 3´no transcritas, por lo que no ofrecen una cobertura óptima de las regiones seleccionadas.

Aunque las aproximaciones mediante estudios de genes candidatos han sido extensamente utilizadas, las conclusiones que se han podido extraer de estos estudios sobre la base genética de las enfermedades complejas son escasas. Tamaños muestrales inadecuados, p valores liberales, falta de medidas de control de calidad adecuadas (como por ejemplo correcciones por tests múltiples) y las restricciones de la hipótesis funcional han resultado en una notable imposibilidad para replicar resultados positivos. Los estudios de asociación caso-control experimentaron un cambio radical cuando aparecieron los primeros *arrays* para el estudio del genoma completo. Este tipo de protocolos no basados en hipótesis previas han resultado muy ventajosos, identificando 14 nuevos loci de susceptibilidad al CCR.

En este contexto, realizamos un estudio GWAS en 881 casos y 667 controles de población española. Nuestro estudio ha conseguido identificar una nueva variante de susceptibilidad localizada en el brazo corto del cromosoma 8 (8p12, SNP rs11987193). La importancia de este hallazgo está además acompañada por el hecho de que éste es también el primer GWAS en CCR que se realiza en una población del sur de Europa. Durante el análisis de asociación, detectamos 64 variantes en 24 loci genómicos que presentaban evidencias de asociación tras el control de calidad y los análisis en la primera fase. Ninguna de estas variantes pudo ser replicada mediante inspección directa de las regiones adyacentes en la cohorte británica CORGI. A pesar de que en un principio esta falta de replicación podría parecer resultado de errores de tipo I durante los análisis estadísticos, encontramos evidencias considerables de que al menos algunos

de estos loci se comportan de forma diferente en EPICOLON y CORGI. Consecuentemente, decidimos genotipar los 10 SNPs divergentes en una segunda cohorte de muestras españolas.

Uno de los marcadores, rs11987193 en 8p12, resultó asociado en esta segunda fase. Aunque los p valores de la réplica no eran extremadamente buenos (p=0.039; OR=0.847 (0.724-0.991)), el análisis en conjunto de las muestras de ambas fases era consistente con la asunción de que el alelo T de este SNP presenta un efecto protector frente al riesgo de desarrollar CCR ($p=2.061x10^{-5}$; OR=0.788 (0.706-0.879)). El SNP se localiza 128kb *downstream* del gen *DUSP4*, que codifica para una fosfatasa de MAP kinasas que ha sido implicada en procesos carcinogénicos. Los miembros de esta familia de proteinas son los responsables de la inactivación de MAPK, previniendo la division cellular y activando los mecanismos de apoptosis inducida por TFGß.

Las particulares cualidades de esta señal de asociación, particularmente la divergencia de MAFs entre EPICOLON y CORGI, podrían ser el reflejo de la variación en los patrones locales de LD del locus que resulta en diferencias en la habilidad de estas poblaciones de capturar la señal de la verdadera variante causal. Esto podría explicar la falta de réplica en la cohorte CORGI, aún cuando las características del SNP (tanto MAF como OR) son similares a las identificadas en el grupo de muestras británicas. Si consideramos esta posibilidad, entonces el estudio de otras poblaciones del sur de Europa podría constituir una herramienta fundamental para el descubrimiento de nuevas variantes de CCR en enfermedades comunes. Además de la nueva variante identificada, es de resaltar que nuestro estudio ha replicado las señales en 6 de los 14 loci descritos hasta ahora (8q24, 10p14, 11q23, 12q13, 15q13 y 18q21). Además, los ORs para el resto de loci eran concordantes con la bibliografía.

Aunque los SNPs son las formas más frecuentes de polimorfismos genéticos, la variabilidad genómica existe en un gran número de formas, desde cambios en una única base hasta grandes reorganizaciones cromosómicas. Por esto, es sensato pensar que además de los SNPs, otras formas de variación también puedan estar implicadas en la modulación de la susceptibilidad genética al CCR. En los últimos años, el descubrimiento de la presencia y abundancia de los CNVs ha acaparado la atención de los estudios de asociación hacia esta forma de marcadores genéticos. A pesar de que varios estudios han descrito que la mayoría de CNVs comunes están intrínsecamente relacionados con los SNPs, existen evidencias de que esta afirmación puede estar sesgada. Por ello, es recomendable que los estudios de asociación analicen directamente los CNVs.

Para minimizar el riesgo de falsos positivos en nuestro análisis, la asignación de los números de copia en cada locus se hizo a través de dos algoritmos diferentes: Birdseye y QuantiSNP. De este modo, los eventos CNV detectados mediante las dos metodologías tendrían menos probabilidades de ser espurios. El grado de solapamiento entre los dos algoritmos para CNVs polimórficos (aquellos con una frecuencia superior al 5% en la población de controles, también denominados CNPs), resultó ser muy elevado. Once de estos CNPs (2p22.3, 4p16.1, 6q14.1, 11q11, 15p11.1-q11.1, 15q13.3, 16p11.2-p11.1, 17q12, 17q21.31 y 18q12.2) además aparecieron como asociados tras los tests de asociación. Es destacable el hecho de que para muchos de estos CNVs, no existen buenos SNP taggers descritos en los arrays comerciales disponibles o incluso en HapMap. Esto apoya nuestra teoría de que la variabilidad en CNVs debe ser inspeccionada por sí misma, y no indirectamente a través de SNPs. Aunque las evidencias de asociación son sólidas para estos 11 CNPs, los p valores para la mayoría son moderados. Por ello, es importante que tanto la presencia de estos CNPs como el

estatus de número de copia sean confirmados por metodologías alternativas, como el MLPA o la qPCR. También resaltamos el hecho de que si esperamos que los CNPs se comporten de igual forma que los SNPs descritos hasta ahora, probablemente se necesiten cohortes con mayores números muestrales para detectar tan sutiles efectos.

En cuanto al segundo objetivo sobre farmacogenética en CCR, se sabe desde hace años que existe una gran variabilidad en la respuesta a fármacos. Tras la administración de cualquier tratamiento se espera que un porcentaje de los individuos no respondan, mientras que otra fracción desarrollará respuestas tóxicas (ADRs, del inglés *Adverse Drug Reactions*). Considerando que la distribución de esta variabilidad se aproxima a la normalidad, es apropiado pensar que las variantes genéticas comunes puedan ser relevantes en la determinación de estos fenotipos. El conocimiento de estos factores permitiría la individualización de los tratamientos mediante la optimización de las dosis farmacológicas y la minimización de los efectos tóxicos. En este contexto, realizamos un GWAS en 221 pacientes de CCR que habían recibido los agentes quimioterapéuticos 5-FU y FOLFOX. A través del análisis de la variación en forma tanto de SNPs como de CNVs conseguimos identificar 11 variantes (rs16857540, rs2465403, rs10876844, rs10784749, rs670454, rs10158985, rs4128317, rs17626122, rs839533, rs10106310 y rs11080058) que muestran evidencias considerables de asociación con una serie de ADRs en un estudio en dos fases. A pesar de que en la mayoría de las ocasiones los resultados positivos en GWAS no proveen con explicaciones biológicas directas sobre los mecanismos de acción de estas variantes, sí que son capaces de proporcionar un medida quantitativa de cómo estos polimorfismos afectan el riesgo individual de sufrir ADRs, por lo que prodrían constituir un factor de decisión esencial a la hora de programar un tratamiento.

Conclusiones

Las conclusiones que se pueden extraer de este trabajo son las siguientes:

1. A pesar de que no hemos conseguido encontrar evidencias suficientes de la presencia de nuevas variantes de riesgo a través de la estrategia de genes candidatos, el SNP rs954353, en la región *upstream* del gen *CYR61* es un potencial candidato que, a nuestro parecer, no debería ser descartado. Creemos que las restricciones en el número muestral pueden haber sido limitantes en nuestro estudio, y por lo tanto la realización de más análisis es necesaria para concretar la relación entre esta variante y la susceptibilidad al CCR.

2. Nuestro GWAS de SNPs ha identificado exitosamente una nueva variante de baja penetrancia, rs11987193, en 8p12. El alelo T de este SNP presenta un efecto protector contra el riesgo de desarrollar CCR. Este locus contiene al gen *DUSP4*, que ha sido relacionado con procesos carcinogenéticos. Dadas las discrepancias en MAFs entre EPICOLON y CORGI, es probable que ambas poblaciones difieran en su habilidad de capturar la señal de la variante real.

3. Hemos identificado 11 CNPs que presentan evidencias sustanciales de asociación con CCR. A pesar de esto, será necesaria la replicación del estatus de número de copia por otras técnicas moleculares, y la replicación en cohortes independientes para determinar por completo la relación de estos CNPs con el riesgo de CCR.

4. Aún cuando estudios previos han descrito que la variabilidad en los CNVs comunes puede ser bien capturada a través de los SNPs, nuestros resultados sugieren que esta

aserción puede estar sesgada por la cobertura incompleta en los diseños de *arrays* previos. Por ello, recomendamos que los estudios de asociación de CNVs se realicen de una forma directa, con el fin de determinar la extensión de la implicación de este tipo de variantes en la modulación del riesgo a CCR.

5. El GWAS para estudiar la susceptibilidad genética a varios ADRs tras tratamientos con 5-FU y FOLFOX ha identificado 11 nuevas variantes que presentan señales de asociación con fenotipos tóxicos. Consideramos pues que los GWAS son una estrategia válida para la detección de variantes farmacogenéticas, e invitamos a la realización de trabajos futuros en este ámbito.

APPENDIX

List of publications:

Pharmacogenomics in Colorectal Cancer: A Genome-Wide Association Study to predict toxicity after 5-Fluorouracil or FOLFOX administration. Fernandez-Rozadilla C, Cazier JB, Crous M, Guinó E, Moreno V, Durán G, Lamas MJ, Paré L, Baiget M, Páez D, López JL, Cortejoso L, García MI, Bujanda L, González D, Gonzalo V, Rodrigo L, Reñé JM, Jover R, Brea-Fernández A, Andreu M, Bessa X, Llor X, Palles C, Tomlinson I, Castells A, Castellví-Bel, Carracedo A, Ruiz-Ponte C for the EPICOLON consortium. *Submitted to the Pharmacogenomics Journal.*

A Two-Phase Case-Control Study for Colorectal Cancer Genetic Susceptibility: Candidate Genes from Chromosomal Regions 9q22 and 3q22. Anna Abulí, Ceres Fernández-Rozadilla, María Dolores Giráldez, Jenifer Muñoz, Victoria Gonzalo, Xavier Bessa, Luisa de Castro, Luis Bujanda, Josep M. Reñé, Angel Lanas, Ana M. García, Joan Saló, Lídia Argüello, Àngels Vilella, Ramiro Carreño, Rodrigo Jover, Xavier Llor, Luis Carvajal-Carmona, Ian PM. Tomlinson, David J. Kerr, Richard S. Houlston, Josep M. Piqué, Angel Carracedo, Antoni Castells, Montserrat Andreu, Clara Ruiz-Ponte, and Sergi Castellví-Bel, for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association. *Submitted to Cancer Epidemiology Biomarkers and Prevention.*

A two-phase case-control study for Colorectal Cancer genetic susceptibility: historical variants and mucins. Anna Abulí, Ceres Fernández-Rozadilla, Virginia Alonso, María Dolores Giráldez, Jenifer Muñoz, Xavier Bessa[2] Rodrigo Jover, Xavier Llor[5], Luis Carvajal-Carmona, Victor Moreno, Angel Carracedo, Antoni Castells, Montserrat Andreu, Clara Ruiz-Ponte, Sergi Castellví-Bel, for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association. *Submitted to BMC Cancer.*

An Update of In Silico Tools for the Prediction of Pathogenesis in Missense Variants. Brea-Fernández AJ, Ferro M, Fernández-Rozadilla C, Blanco A, Fachal L, Santamariña M, Vega A, Pazos A, Carracedo A and Ruiz-Ponte C. Current Bioinformatics. *In press.*

The GREM1, BMP4 and BMP2 loci harbor multiple common susceptibility variants for colorectal cancer. Luis G Carvajal-Carmona, Sara E Dobbins, Albert Tenesa, Angela M Jones, Kimberley Howarth, Claire Palles, Peter Broderick, Emma EM Jaeger, Susan Farrington, Annabelle Lewis, James GD Prendergast, Alan M Pittman, Evi Theodoratou, Bianca Olver, Rebecca A Barnetson, Steven Penegar, Ella Barclay, Nicola Whiffin, Lynn Martin, Amy Lloyd, Maggie Gorman, The COGENT Consortium, The CORGI collaborators, Clara Ruiz-Ponte, Ceres Fernandez-Rozadilla, Antoni Castells, Angel Carracedo, Sergi Castellvi-Bel, David Duggan, David Conti, Jean-Baptiste Cazier, David J Kerr, Harry Campbell, Oliver Sieber, Lara Lipton, Peter Gibbs, Grant Montgomery, Joanne Young, Paul Baird, Brent Zanke, Steven Gallinger, Polly Newcomb, John Hopper, Mark A Jenkins, Lauri A Aaltonen, Jeremy Cheadle, Paul Pharoah, Graham Casey, Malcolm G Dunlop, Ian PM Tomlinson, Richard S Houlston. PLOS Genetics. *In press.*

Single nucleotide polymorphisms in the Wnt and BMP pathways and colorectal cancer risk in a Spanish cohort. Ceres Fernández-Rozadilla, Juan Clofent, Luisa de Castro, Alejandro Brea-Fernández, Xavier Bessa, Anna Abulí, Montserrat Andreu, Rodrigo Jover, Xavier Llor, Antoni Castells, Sergi Castellví-Bel, Angel Carracedo, Clara Ruiz-Ponte* for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association. PLoS One. Sep 2010.

Nuevos métodos de diagnóstico molecular: genómica. Clara Ruíz-Ponte, Ceres Fernández-Rozadilla. GH continuada. Jul-Aug 2010.

Colorectal cancer susceptibility QTLs in mice as a novel approach to detect low-penetrance variants in humans: a two-stage case-control study. Ceres Fernández-Rozadilla, Rosa Tarrío, Juan Clofent, Luisa de Castro, Alejandro Brea-Fernández, Xavier Bessa, Anna Abulí, Montserrat Andreu, Rodrigo Jover, Rosa Xicola, Xavier Llor, Antoni Castells, Sergi Castellví-Bel, Angel Carracedo, Clara Ruiz-Ponte[*] for the Gastrointestinal Oncology Group of the Spanish Gastroenterological Association. Cancer Epidemiology Biomarkers and Prevention. Feb 2010.

Molecular analysis of the *APC* and *MYH* genes in Galician and Catalonian FAP families: a different spectrum of mutations? Nuria Gómez-Fernández, Sergi Castellví-Bel, Ceres Fernández-Rozadilla, Francesc Balaguer, Jenifer Muñoz, Irene Madrigal, Montserrat Milà, Begoña Graña, Ana Vega, Antoni Castells, Angel Carracedo, Clara Ruiz-Ponte[§]BMC Medical Genetics. Jun 2009.