Studies of inherited predisposition to cancer

Richard S Houlston

Institute of Cancer Research,
SM2 5NG

Publications submitted for degree of DSc in accordance with the regulations of Imperial College, London 2012
Statement on how my publications have contributed to the advancement of knowledge and learning in the field

EARLY RESEARCH

The research studies undertaken for my MD thesis were concerned with the impact of genetic variation in the apolipoprotein B (Apo-B) and Apo-E genes on lipoprotein metabolism. I demonstrated that the catabolic rate of low density lipoprotein (LDL) is influenced by variation in the Apo-B gene and this directly impacts on circulating levels of LDL; Notably that polymorphic variation in Apo-B is a determinant of LDL levels in the general population and that germline mutation of the gene can result in a clinical phenotype analogous to familial hypercholesterolaemia.

This period of research was followed by a Clinical Fellowship from the Imperial Cancer Research Fund during which I worked on the genetic epidemiology of breast, ovarian and colorectal cancer. Using large datasets I calculated age-specific familial cancer risks for these cancers and using statistical modelling determined the most likely genetic models of inherited predisposition to breast, ovarian and colorectal cancer (CRC). Using these data a screening programme for the relatives of patients CRC based on their calculated lifetime risk was developed and implemented as part of the North East Thames Regional Clinical Genetics Service. It was demonstrated that family history of cancer could be used to identify those at risk of colonic cancer and to target appropriate screening. In the CRC family history clinic colonoscopic surveillance was shown to detect a high number of premalignant colonic polyps in those calculated to be at high genetic risk. An analogous programme for familial breast cancer was subsequently implemented for individuals with a family history of breast cancer. In this clinical setting use of familial history was shown to be an effective means of identifying women at risk of breast cancer who might benefit from early mammographic surveillance.

The findings and experience demonstrated the clinical utility of dedicated family cancer clinics and established a framework for their operation within the United Kingdom. Collectively work undertaken during this period formed the basis of a PhD thesis.

CURRENT RESEARCH INTERESTS AND ACHIEVEMENTS

I am currently Professor of Molecular and Population Genetics, Division of Genetics and Epidemiology at the Institute of Cancer Research. The central theme of my research is to understand the biological basis of inherited susceptibility to cancer and specifically the identification and characterisation of genetic variation influencing cancer risk.

Over the past 40 years the study of inherited susceptibility to cancer has proved to be a most informative area of cancer research – and it continues to be so. The results of this research, namely the identification of new susceptibility genes, provides for an understanding of the mechanisms of tumour biology, offering potential targets for novel therapeutic interventions. The ability to identify those at increased risk is of clinical relevance, in terms of primary and secondary interventions. Finally given the difficulties in unambiguously identifying causative exposures for some cancers, genetic associations are likely to be increasingly valuable via the
functional links they reveal that either endorse current aetiological hypotheses or suggest new ones that merit testing via gene-environment specific hypotheses.

IDENTIFICATION OF DISEASE GENES THROUGH POSITIONAL CLONING

My first work in the area of identification of cancer genes was on Juvenile Polyposis (JPS), a rare classical Mendelian dominant condition conferring a high risk of CRC. We undertook a large family collection based on strict ascertainment criteria, which ultimately through international collaboration led to the identification of the disease gene, SMAD4, for JPS. Its identification has allowed families to be tested for JPS in the diagnostic and predictive settings in routine clinical practise.

A similar emphasis on detailed family ascertainment allowed us to conduct a genome-wide linkage scan of families segregating hereditary leiomyomatosis and renal cancer (HLRCC), and localise the gene for HLRCC to 1q42.3-q43. Subsequently, as part of an international collaboration, through positional cloning we identified germline mutations in the gene encoding the Krebs cycle enzyme fumarate hydratase (FH) as the basis of HLRCC. While HLRCC is a rare disease the demonstration that mutation of FH can be a basis for tumour development provided contributing evidence of pseudo-hypoxic drive in the development of cancer.

Subsequent to these studies we were one of the first research groups to implement high-density SNP arrays to conduct genome-wide linkage scans. While the denseness of SNP marker maps means that there is increased power and improved localisation of disease genes the high density of SNP maps leads to the problem of linkage disequilibrium (LD) between markers inflating linkage statistics. To address this issue we developed the program SNPLINK to incorporate LD between markers into the likelihood calculations so that expected haplotype frequencies are correctly estimated.

Prior to performing linkage scans of CRC and chronic lymphocytic leukaemia (CLL) families to search for novel disease loci we extensively piloted the utility of using high-density SNP arrays. Firstly, we evaluated a pre-release version of the Affymetrix 10K array successfully using this technology to localise and identify PTF1A as a cause of recessive cerebellar and pancreatic agenesis. Secondly, we validated the use of production Affymetric 10K arrays identifying disease genes for a number of Mendelian disorders including recessive paroxysmal nocturnal hemoglobinuria, vasculitis, merosin-positive congenital muscular dystrophy, craniosynostosis and dominantly inherited renal dysplasia.

Following on from these studies we analysed 206 CLL families identifying a disease locus at 2q21.2, thereby providing evidence for high-moderate penetrance susceptibility to CLL. Through mutation screening we subsequently demonstrated that germline mutation of the gene CRCX4, which maps to 2q21.2 is implicated in a subset of familial CLL.

To investigate Mendelian susceptibility to CRC we conducted a genome-wide linkage screen of 104 CRC families in which germline mutations in the then known CRC predisposition genes had been excluded. The strongest evidence of linkage was provided by loci at 9p, 3p24, 13q31 and 17q24, albeit non-significant at a genome-wide threshold. These data indicate that if additional high-moderate penetrance susceptibility genes exist for CRC, each accounts for only a small fraction of the familial risk.
IDENTIFICATION OF COMMON RISK VARIANTS FOR CANCER

Over the last decade the opportunities for identifying high-risk cancer genes through classical linkage and positional cloning strategies have dwelled. This has sparked renewed interest in the notion that much of the inherited predisposition to cancer is mediated through various combinations of common and rare-disease causing genetic variants. Anticipating the requirement for large datasets of cancer cases to search for these classes of cancer susceptibility gene we developed the necessary infrastructure and protocols for large-scale sample collection together with establishment of international consortia. These initiatives have allowed us to generate world-class biobanks for multiple cancer types.

Initially our search for common genetic variants affecting cancer risk was largely conducted adopting a candidate gene approach making use of analytical platforms technologies for parallel processing of genotyping. The spectrum of mutations in Mendelian disease genes, coupled with issues of statistical power, provided us with a rationale for pursuing association analyses targeting non-synonymous SNPs (nsSNPs). To increase the likelihood of identifying predisposition variants we classified and catalogued the predicted functionality of all nsSNPs catalogued by dbSNP in genes relevant to the biology of cancer through cross referencing of KEG and other databases. In addition to studies of CRC, breast cancer, CLL our study of lung cancer provided evidence for a role for variants of the IGF and BAT3 in disease aetiology.

Subsequent to studies focused on candidate genes using our infrastructure, we has been at the vanguard of implementing genome-wide association studies (GWAS) to identify common, low-penetrance loci for cancer without prior knowledge of location and function. We has successfully led GWAS studies of CRC, lung cancer, glioma, meningioma, CLL, acute lymphoblastic leukaemia (ALL), Hodgkin’s lymphoma (HL) and multiple myeloma (MM) being the first research group to identifying novel common disease variants for each malignancy. Importantly, prior to our studies no specific risk factors for ALL, CLL and MM had been identified. Collectively our work in this field of research has so far resulted in the identification of over 50 novel cancer loci constituting a large fraction (perhaps ~25%) of all common variants thus far identified by all researchers worldwide. In the design of GWAS we have shown in our studies of CRC and CLL the theoretical value of using cases enriched for genetic susceptibility by virtue of family history as a means of increasing study power to identify novel disease causing variants through association-based analyses.

As well as vindicating the hypothesis of low penetrance susceptibility to cancer our findings have provided insight into the biological basis of tumourigenesis, emphasising the role of genetic variation in developmental genes as determinants of cancer susceptibility for a number of malignancies.

One of the anticipated deliverables from GWAS studies was that the identification of risk variants would provide fresh insights into cancer biology. Indeed few of the genes implicated by the GWAS scans have previously been evaluated in targeted association studies. Through the work of our group, insights into new pathways of tumourigenesis are emerging; for example, we have demonstrated that genetic variation in the B-cell developmental genes, IKZF1, CEBPE, IRF4, IRF8, SP120, BAK1, and ARID5B, has been shown to determine the risk of B-cell tumours.
Furthermore we have shown many of the risk variants have cis-acting effects on gene expression rather than influencing risk through sequence variation in the expressed proteins.

Similarly, several of the 20 CRC loci we have identified provide strong evidence for the involvement of components in the transforming growth factor-β (TGFβ) superfamily signalling pathway in CRC development, most notably BMP2, BMP4, SMAD7 and GREM1. Intriguingly, we have demonstrated an association between chromosome Xp22.2 encompassing the developmental gene SHROOM2 and CRC risk. This represents the first evidence for a role of variation in the X-chromosome variation in predisposition to a non-sex specific cancer characterised by different incidence rates in men and women.

While the well established HLA association with HL represents a very strong genetic effect, the identification of risk variants at 2p16.1, 8q24.21 and 10q14 we have made has implicated important roles for networks involving MYC, GATA3 and the NFκB pathway in HL disease aetiology. Our findings in glioma have also highlighted the importance of variation in genes encoding components of the CDKN2A-CDK4 signalling pathway in tumour development. Moreover, this pathway, elucidated through the extended interaction network of CDKN2A, incorporates TERT (through mutual interaction with HSP90) and other genes (including CCDC26) identified as risk factors. Gliomas a heterogeneous and following on from initial gene discovery experiments we have shown that the risk variants at 5p15.33 (TERT), 8q24.21 (CCDC26), and 20q13.33 (RTEL) have subtype-specific effects consistent with different aetiological basis to the various glioma histologies.

Similarly our work on the other major primary brain tumour, meningioma has implicated dysfunctional Wnt signalling as a biological basis for tumour development by virtue of variation in MLLT10. This together with our previous observation of a relationship between BRIP1 variation and meningioma risk represent the only robust genetic associations for this tumour type reported thus far.

Most recently the GWAS of MM we performed has identified three loci influencing MM risk. Our observation that genetic variation affecting CDCA7L and ULK4 genes affects MM risk has for the first time directly implicated genetically determined deregulation of MYC and mTOR-mediated autophagy in the aetiology of the disease.

Importantly our studies have provided the first direct and robust evidence for inherited genetic susceptibility to ALL, CLL, MM and lung cancer. This was especially profound for lung cancer a malignancy frequently cited as a solely attributable to environmental exposure. While it has long been postulated that individuals may differ in their genetic susceptibility to develop lung cancer in response to genotoxic insult prior to our studies evidence for such an assertion has been lacking. Through our GWAS of lung cancer we initially identified risk loci for lung cancer at 5p15.33 (TERT/CLPTM1), 6p21.33, and two at 15q25.1 (CHRNAS-CHRNA3-CHRNAS). These data thus provided the first evidence for common genetic susceptibility to lung cancer. We have subsequently shown through collaborative pooled analyses of other GWAS datasets that variation in RAD52C and CDKN2A are risk factors for squamous lung cancer and the 5p15.33 (TERT) association is specific for risk of lung adenocarcinoma.

Most SNP associations identified to date have been tumour specific, which is consistent with the epidemiological observations that most familial cancer risks are tumour specific. Evidence for
pleiotropic effects (reflecting generic effects or lineage-specific effects) has been provided by our work showing that the 5p15.33 (TERT–CLPML1) locus influences the risk of many tumour types including glioma and lung cancer and the locus at 9p21.3 (CDKN2A–CDKN2B) influences the risk of both glioma and ALL risk in addition to melanoma.

In contrast to high-penetrance susceptibility, the effect of low-penetrance variants on tumour phenotype may be limited or absent. Exploration of the relationship between SNP genotypes and tumour phenotypes is still in its infancy, but some associations are becoming apparent. One of the most striking genotype–phenotype relationships identified to date is the 10q21.2 (ARID5B) ALL association, which seems to be highly selective for the subset of B-cell precursor ALL with hyperdiploidy directly implicating the developmental gene ARID5B in disease phenotype.

Given the difficulties in unambiguously identifying causative exposures for many cancers, genetic associations have the potential to endorse current aetiological hypotheses or suggest new ones that merit testing through gene- and/or environment-specific hypotheses. Demonstration of an effect on cancer risk that is mediated by an environmental exposure has been provided by the nicotinic acid receptor (CHRNA3–CHRNAS) locus that is associated with lung cancer through genetic variation influencing an individual’s propensity to smoke. Since the 15q24 association signal is a consequence of the D398N substitution in CHRNA5 which causes decreased response to nicotine agonists there is evidence for genetically-determined nicotine dependency with biological plausibility.

Our findings that variation in developmental genes is commonly the basis of GWAS signals for cancer generally favours such variation impacting early in tumourgenesis rather than impacting on late disease expression. Testing for association between risk variants and progenitor lesions theoretically offers the means of examining such an assertion. For most tumours this is unfortunately not possible. For CLL this is possible since we have shown that monoclonal CLL-phenotype cells (monoclonal B-cell lymphocytosis, MBL) detectable in ~3% of otherwise healthy persons and represent the progenitor lesion for CLL. Our observation that the risk variants for CLL influence affect the risk of MBL is consistent with the variants impact on early stage development of CLL rather than disease progression per se.

Modelling all the single nucleotide polymorphisms (SNPs) in a GWAS simultaneously provides a means of deriving an unbiased estimate of the heritability explained by common variation. In concert with our gene discovery efforts using GWAS-based strategies we have applied this form of statistical modelling to ALL and CLL GWAS datasets to derive estimates for heritability. Through these analyses we have shown that 59% of the total variation in CLL risk can be accounted for by common genetic variation, thereby providing the first direct data for a polygenic basis for susceptibility to CLL. Prior to our GWAS of ALL evidence although speculated was distinctly lacking and much of the contemporary thinking was that, given indirect evidence for the role of infection in disease aetiology any genetic susceptibility would be mediated through disordered HLA-associated. Our analyses have quashed this hypothesis but have shown that 28% of the variance in childhood ALL risk can be ascribed to common variation.
Genetic and functional basis of GWAS signals

To investigate if associations may have arisen owing to independent correlation of a tagSNP with more than one functional SNP, we searched for novel CRC susceptibility variants close to the BMP-pathway genes GREM1, BMP4, and BMP2. We have shown that independent CRC predisposition SNPs close to BMP4 and BMP2. Near GREM1 we also found using fine-mapping that the previously-identified association between tagSNP rs4779584 and CRC was a consequence of two independent signals. As exemplified by MLH1-93G>A polymorphism there is increasing evidence that high- and low-penetrance variants can map to the same gene. Further support for such an assertion is supported by our observation of inactivating BMP4 mutations as a basis of familial CRC. Our work in this area has thus served to emphasise that genetic fine-mapping studies can deconvolute associations, thus explaining some of the apparently missing heritability of common diseases.

A long term outcome of GWAS is that knowledge gained about the underlying molecular basis of CRC may lead to the development of innovative therapeutic and preventative measures. Many of the risk loci identified thus far by GWAS map to non-coding regions of the genome. For example, the 8q24.21 region is one of the most intriguing and important loci to emerge from GWASs. The genomic interval harbours independent loci with different tumour specificities including ones for CRC, CLL and HL. The region to which these cancer associations map is, however, bereft of genes or protein-coding transcripts. Identification of the causal basis of association signals identified through GWAS is challenging and we has played a major role in elucidating the functional basis for the CRC associations, work that has underscored the role of inherited differential gene expression as a determinant of cancer susceptibility rather that mutation impacting on protein sequence which has typified Mendelian cancer susceptibility.

Using a combinatorial approach of bioinformatics and molecular studies we have demonstrated the genomic region harbouring the 8q23 variant for CRC influences eukaryotic translation initiation factor 3, subunit H (EIF3H) expression and that up-expression of EIF3H gene increases CRC growth and invasiveness thus providing a biological mechanism for the 8q23.3 association. Similarly our work has demonstrated cis-acting regulation of the TGF-beta signalling gene SMAD7 as a basis for the 18q24 association and cis-acting regulation of BMP4 provides a basis for the 14q22.2 association.

At 8q24.21 we showed rs6983267 was maximally associated with CRC risk. A collaboration with Lauri Aaltonen (Helsinki University) led to the demonstration that rs6983267 annotates an enhancer element affecting the binding of Wnt-regulated transcription factor TCF4. This coupled with the finding that rs6983267 directly interacts with the MYC promoter has provided a biological basis for the 8q24 association.
GENETIC EPIDEMIOLOGICAL ANALYSES

Studies of CLL

Although familial clustering of CLL had long been recognised direct evidence for inherited genetic susceptibility has been lacking. We have made a significant contribution to establishing that inherited genetic factors play a role in the development of CLL. Even prior to our linkage and association studies through the ascertainment of striking families segregating CLL we provided overwhelming case for the existence of genetic predisposition. Furthermore, our observations on the repertoire and frequency of IGVH usage in familial and sporadic CLL favour a genetic basis to CLL development rather than a simple environmental aetiology. Coupled with such studies our studies of MBL have provided insight in disease development.

Risk prediction and statistical modelling of familial risk

The distribution and genotypic risk conferred by disease alleles is crucial for determining the applicability to relative or absolute risk circumstances. Diagnostic testing for highly penetrant mutations is now part of standard clinical care for cancer families in many countries. The identification of individuals at increased risk allows the targeting of cancer prevention strategies and can increasingly influence cancer treatment.

Our work on the estimation of familial cancer risks has been largely focused on those relating to CRC. In addition to deriving age-specific familial CRC risks we have stratified risks by molecular features in cancers. These analyses indicate that the majority of the familial CRC risk associated with micro satellite unstable CRC is a consequence of germline MMR mutation/variation. However, ~70% of the familial CRC risk current remains unexplained. Statistical modelling of this “missing heritability” was shown to be compatible with polygenic/recessive susceptibility.

Using the largest dataset to date we have derived age-specific CRC risk associated with mutations in the base-excision repair gene MUTYH. We have shown that while biallelic mutations are associated with a high CRC risk penetrance is incomplete at age 60. We have also shown MUTYH mutation screening should be directed to patients with APC-negative polyposis and early-onset proximal microsatellite stable CRC an expanded clinical phenotype needs be recognized. This information is directly relevance to clinical counselling.

There has been limited data on the spectrum and risk for cancer associated with germline serine/threonine protein kinase 11 (STK11) mutations that cause Peutz-Jeghers syndrome (PJS). To address this deficiency we analyzed the incidence of cancer in 240 individuals with PJS possessing germline mutations in STK11, the largest study of its type to date. The most common cancers represented were gastrointestinal in origin-gastroesophageal, small bowel, colorectal, and pancreatic. In women, the risk for breast cancer was substantially increased. Our analysis showed similar cancer risks between missense and truncating mutation carriers. These results quantitatively showed the spectrum of cancer risk associated with STK11 germline mutations in the context of PJS and provide the most comprehensive data for defining surveillance regimens thus far.
Public health value of common variants on cancer risk

At present, the power of models that incorporate all known common risk alleles for individual-level risk predication is limited, although there is clearly potential for this to improve substantially as more variants are found. This may in turn have important health implications for the provision of cancer screening; for example, in determining who should undergo colonoscopy. Nevertheless we have demonstrated by modelling the public health potential of risk profiling using composites of risk variants for CRC that stratification of the population into CRC risk categories is feasible, informing targeted prevention and surveillance within the population to be optimally configured.
CONTRIBUTION TO MULTI-AUTHORED PUBLICATIONS

A. PUBLICATIONS CONSIDERED BY TOPIC

STUDIES OF GENETIC VARIATION IN LIPOPROTEIN GENES AND RELATED STUDIES
In my MD studies, under the supervision of Humphries, I performed all molecular analyses [1, 2, 7, 8, 9, 10, 13, 14, 16, 17, 20, 39]. During this period I assayed Lp(a) to investigate its role as a risk factor for ischaemic heart disease in the context of familial hypercholesterolaemia [6] and participated in clinical trials of studies to assess the effectiveness of lipid lowering strategies [3,4].

ESTIMATION OF FAMILIAL CANCER RISKS AND STATISTICAL MODELLING OF FAMILIAL CANCER
For my PhD studies, under the supervision of Slack and Morton, I calculated the familial risks for breast, ovarian and colorectal cancer (CRC) and determined the most probable genetic basis of familial aggregation using segregation analysis [15, 18, 19, 21, 33]. Coupled with these studies and as part of the work towards my thesis I evaluated the value of using familial cancer risks in clinical practice in determining the screening requirements of patients with family histories of CRC and breast cancer [11, 23, 31]. During this period I supported studies of the familial aggregation of melanoma by Rustin through case ascertainment and statistical analysis of cancer risks [22, 29]. Following on from my PhD I either performed myself or supervised the statistical analysis of other studies familial cancer risks [34, 45, 81, 91, 107, 128, 165, 177, 214, 266, 267]. For studies of CRC risks stratified by MSI or MUTHY mutation status I supervised both statistical and molecular analyses [266, 267].

I have also conceived and supervised a study of the impact of common variation on CRC risk [] and co-led a similar multicentre analysis [].

Risks and phenotype associated with CRC SNPs [353]

IDENTIFICATION OF DISEASE GENES THROUGH POSITIONAL CLONING-BASED STRATEGIES
The first work of my research group in disease gene identification was through participating in multicentre studies of juvenile polyposis. We performed family ascertainment, linkage and mutation analyses [56, 58, 63, 73, 83]. Following on from this study, in collaboration with Tomlinson, we performed a genome-wide linkage scan of HLRCC [100] and through multicentre collaboration, involving the research groups of Tomlinson and Aaltonen, we performed fine mapping and mutational analyses identifying FH as the disease gene for HLRCC [109]. I subsequently supervised the search for somatic mutation of FH in sporadic leiomyomas and other tumours [113, 127, 168].

In addition to conducting genome-wide linkage scans my research group evaluated linkage at specific putative loci for a number of diseases including familial non-medullary thyroid cancer [47, 64, 65, 87, 94, 181].

We successfully implemented the use of SNP arrays for performing linkage analyses, rather than relying on microsatellites, developing in house software to deal with issues of linkage disequilibrium between SNPs [140, 155, 156, 162, 170, 173, 174, 184, 187, 218, 256].
My research group, under my supervision, has also conducted linkage analyses as part of external collaborations [48, 49, 51, 52, 53, 54, 57, 60, 75, 82, 97, 98, 181, 281, 294]. Following on from our demonstration of an association between common variation at BMP4 and CRC I led an in-house study which demonstrated mutation of BMP4 as a cause of high-penetrance CRC susceptibility [319]. Using a similar candidate based analysis of familial cases I supervised the mutational analysis of NPAT as a cause of Hodgkin’s disease in a study led by the Finish group of Aaltonen[340].

I was one of the investigators who established the GLIogene consortium in order to study familial glioma [222]. We have undertaken genotyping as part of this international effort [296, 310, 346, 354, 369, 376, 379] and have led a study of the role of TP53 and P16 mutation and allergy genes in disease aetiology [300, 307].

IDENTIFICATION OF CANCER SUSCEPTIBILITY LOCI THROUGH ASSOCIATION-BASED ANALYSES

**Candidate-gene studies**

To search for low penetrance variants influencing cancer risk the work of my research group was initially directed towards evaluation of variants in candidate genes genotyping of through case-control series. These studies were conceived by me and I supervised all laboratory and statistical analyses [74, 92, 137, 142, 146,147,159, 178, 182, 186, 188, 189, 190, 194, 198, 200, 201, 203, 209, 211, 215, 216, 220, 223, 217, 225, 231, 234, 235, 236, 237, 243, 247, 248, 250, 251, 260, 264, 271, 274, 283, 307, 339]. As part of collaborative studies in which I was not lead author I either performed statistical analyses to support the study or supervised statistical/molecular analyses [90, 108, 111, 116, 118, 121, 124, 136, 139, 148, 149, 151, 158, 172, 175, 191, 208, 213, 240, 239, 255, 257, 259, 263, 268, 272, 278, 285, 291, 296, 297, 302, 303, 308, 310, 311, 312, 316, 317, 323, 326, 330, 331,343, 342, 371, 373, 375, 377 , 362, 364, 366]. I also either solely performed or jointly performed/directed meta-analyses of cancer risks associated with specific polymorphic variants [67, 69, 85, 86, 101, 131, 306]; and performed meta-analyses as part of projects led by other researchers [205, 269].

To support my association studies I have conceived, led and supervised the development of in-house biobanks and international consortia [226, 249, 334, 355, 362].

**Genome-wide association studies**

I have conceived and led multi-centre genome-wide association studies (GWASs) of acute lymphoblastic leukaemia (ALL) [276, 301] chronic lymphocytic leukaemia (CLL) [254, 292, 309, 337, 368], Hodgkin’s disease (HD) [322], meningioma [347] and jointly conceived and directed GWASs of CRC [202, 227, 230, 238, 262, 287, 320, 367, 375], lung cancer [241, 261, 279, 289, 338, 341, 344 359,377], glioma [275, 336] and multiple myeloma [352]. To elucidate the functional basis of GWAS associations we have fine-mapped the risk loci for CLL [293], led multi-centre fine mapping studies of CRC loci and performed associated in vitro studies [270, 314, 358] and performed statistical analyses to support collaborative efforts [273, 335, 356]. We have exploited the GWAS data to determine: (1) Heritability of ALL [365] and CLL [378] attributable to common variation; (2) Role of MHC variation in ALL [324], HL [342] and CLL [372]; (3) Role of homozygosity as a risk factor for breast and prostate cancer [299]. In addition we contributed to the statistical analysis study of homozygosity and CRC risk performed Tomlinson [280]. Finally, using data from the multicentre GWAS of CRC I supervised the statistical analysis of the
relationship between mitochondrial variants, coding SNPs and risk [264, 271]. Following on from our studies of ALL we have through multicentre collaborations examined the impact of risk variants in non-UK populations [228, 318]; my role being leadership with respect to requisite analyses.

OTHER STUDIES

Studies of familial CLL: I led descriptive studies of familial CLL through review of reported cases and ascertainment of cases in house, in collaboration with Yuille and Catovsky [22,84] which provided evidence for Mendelian inheritance; and collaborated on the systematic follow up of a large Danish family involving over 100 individuals [167]. I performed statistical analysis which provided support for anticipation [59] and pseudo-autosomal inheritance in CLL [78]. I supervised the study of CAG repeats in familial CLL in collaboration with Benzow and Koob who performed molecular studies [119] and collaborated with on CGH analysis of familial CLL performed by Shipley [102]. I supervised ascertainment of CLL families and the mutational analysis of CXCR4 [284]. I devised and led an investigation demonstrating MBL as the progenitor lesion for CLL with Rawstron performing flow studies [112]; subsequently I participated in follow on review of the MBL phenotype [183]. I conceived and led a multicentre study demonstrating the risk SNPs for CLL influence MBL risk [315]. In additional studies of the aetiology of CLL I was a collaborator responsible for performing statistical analyses [191, 223].

Studies of Peutz-Jeghers syndrome (PJS): I conceived and established a multicentre study of PJS cases and families supervising the estimation of PJS-associated cancer risks [133, 150, 195], spectrum of STK11 mutation and genotype-phenotype correlation [164, 176, 196], and a search for a novel disease locus by chromosomal mapping [153].

Studies of breast cancer: I conceived and led the ascertainment of LCIS cases and supervised mutational analysis of E-cadherin [79] and mismatch repair genes [96].

Specific statistical analyses: As part of a collaborative projects I have: (1) Estimated the penetrance of WT1 mutation from family data [81]; (2) Determined the statistical distribution of low level microsatellite instability in CRC [110] and allelotype [125]; (3) Contributed to an understanding of the variation in FAP phenotype by performing segregation analyses [115]; (4) Collaborated, by providing statistical assistance in devising derive optimal strategies for disease gene identification [50], an understanding of missing heritability [321] and strategies for GWAS [325].

Studies of gastric cancer: I conceived and led the ascertainment of gastric cancer cases and supervised mutation analysis of E-cadherin in cases [61].

Studies of Proteus syndrome: Together with Harper I established a biobank of tissues from proteus patients and have supervised the in-house evaluation of PTEN , VG5Q [95, 185], and participated in studies of copy number variation as a cause of disease [305].

Clinical studies: During my general professional and specialist training in medicine I published a number of dysmorphic syndromes and cancer-related clinical phenotypes [26, 27, 28, 32, 35 36, 38, 40, 41, 42, 43, 138, 152], co-authored description of other syndromes and cancer families
[145, 154, 229] and participated in the evaluation of gene testing [304]. In collaboration with Lucassen we undertook a survey of attitudes in clinical cancer genetics [76]. In addition I participated in a number of clinical studies by either performing statistical analyses [5] or by ascertainment highly selected patients /samples for studies [44, 46].

**Pathology and studies of prognostic markers:** I conceived and led studies evaluating methods for detection of mismatch repair defects in CRC [80, 93], an evaluation of inter-observer reporting of CRC histology [228]. Additional I have supervised the analysis of the mutational spectra in a number of cancers [179, 197, 202, 204, 219] and collaborated in the analysis of data in other studies [192].

I conceived and led studies evaluating methods evaluating markers of CRC and lung cancer prognosis involving in-house genotyping/mutational/statistical analysis [206, 207 210, 212, 221, 265].

**Co-morbidity ALL [370]**

I conceived and led the ascertainment of cases and supervised mutation analysis Bcl10 in CRC [62]lfs [71]

Particated in mmr in t-cell prolympchocty leukaemia [70], cll 72

BAT26 [77]

**Studies of celiac disease:** After establishing a cohort of celiac disease families I conceived and supervised Popat in studies evaluating the role of CTLA4 [104, 114], CD28 [104], in disease aetiology. Together with Howdle I conducted a survey of small bowel cancer risk associated with celiac disease, formulating study design and performing statistical analyses [134]. Additionally I supervised mutation testing of tapsin [88].

**CHEK2 soda [117]**

Radiosensitivity:
Sod2 in radio [122], meningioma [333], FRFR2 and HD [356]

**Ocular melanoma**
After establishing a cohort of ocular melanoma patients through collaboration with Damato conceived and supervised studies examining the role of BRCA2, CDKN2A [123] and MC1R [126] in this cancer.

**MUTYH in lung cancer [135]**
Review and other articles

B. PUBLICATIONS CONSIDERED IN CHRONOLOGICAL ORDER

PAPERS

1. Part of my MD studies
2. Part of my MD studies
3. Part of my MD studies
4. Contributor to a clinical trial
5. Performed statistical analyses
6. Study conceived, designed and conducted by me
7. Part of my MD studies
8. Part of my MD studies
9. Part of my MD studies
10. Part of my MD studies
11. Part of my PhD studies
12. Part of my PhD studies
13. Part of my MD studies
14. Part of my MD studies
15. Part of my PhD studies
16. Part of my MD studies
17. Part of my MD studies
18. Part of my PhD studies
19. Part of my PhD studies
20. Part of my MD studies
21. Part of my PhD studies
22. Part of my PhD studies
23. Part of my PhD studies
24. Part of my PhD studies
25. Part of my PhD studies
26. Report of clinical syndrome by me
27. Report of clinical syndrome by me
28. Report of clinical syndrome by me
29. Performed statistical analysis of familial cancer risks
30. Part of my MD thesis
31. Report of clinical syndrome by me
32. Report of clinical syndrome by me
33. Part of my PhD thesis
34. Participation in study design and data interpretation
35. Report of clinical syndrome by me
36. Report of clinical syndrome by me
37. Study conceived, designed and statistical analyses performed by me.
38. Report of clinical syndrome by me
39. Performed genotyping as part of association study
40. Report of clinical syndrome by me
41. Report of clinical syndrome by me
42. Report of clinical syndrome by me
43. Report of clinical syndrome by me
44. Contributed clinical data on family with TP53 mutation
45. Study conceived, designed and supervised by me
46. Clinical collaborator on report of familial tertatoma
47. Study conceived, designed and supervised by me
48. Clinical collaborator
49. Clinical collaborator
50. Jointly conceived study, design and execution
51. Performed statistical analyses
52. Performed statistical analyses
53. Performed statistical analyses and data interpretation
54. Clinical collaborator
55. Clinical collaborator
56. Multi-centre study to which we contributed linkage and mutational data on JPS families
57. Performed statistical analyses and data interpretation
58. Study conceived, designed and supervised by me
59. Contributed to study design, performed statistical analyses
60. Contributed to study design, performed statistical analyses
61. Study conceived, designed and supervised by me
62. Study conceived, designed and supervised by me
63. Study jointly conceived and supervised by me
66. Study conceived, designed and supervised by me
67. Study conceived, designed and supervised by me
68. Study conceived, designed and supervised by me
69. Study solely undertaken by me
70. Contributor to study design and data interpretation
71. Study conceived, designed and supervised by me
72. Study conceived, designed and supervised by me
73. Study conceived, designed and supervised by me
74. Study conceived, designed and supervised by me
75. Performed statistical analyses and data interpretation
76. Study jointly conceived and executed
77. Contributed to study design and data interpretation
78. Study jointly conceived and executed
79. Study conceived, designed and supervised by me
80. Study conceived, designed and supervised by me
81. Performed statistical analyses and data interpretation
82. Performed statistical analyses
83. Multi-centre study to which we contributed mutational data on JPS families
84. Study conceived, designed and supervised by me
85. Study conceived, designed and supervised by me
86. Study conceived, designed and supervised by me
87. Study conceived, designed and supervised by me
88. Contributed molecular data
89. Contributed molecular data
90. Contributed molecular data
91. Study conceived, designed and supervised by me
92. Study conceived, designed and supervised by me
93. Study conceived, designed and supervised by me
94. Study conceived, designed and supervised by me
95. Study jointly conceived, designed and supervised by me
96. Study conceived, designed and supervised by me
97. Performed statistical analyses
98. Performed statistical analyses
99. Contributed molecular data
100. Study conceived, designed and supervised by me
101. Study jointly conceived, designed and executed by me
102. Contributed molecular data
103. Study conceived, designed and supervised by me
104. Study conceived, designed and supervised by me
105. Contributed to design of molecular analyses
106. Study conceived, designed and supervised by me
107. Study conceived, designed and supervised by me
108. Performed statistical analyses and data interpretation
109. Multi-centre study to which we contributed equally
110. Performed statistical analyses and data interpretation
111. Contributed molecular data
112. Study conceived, designed and supervised by me
113. Study conceived, designed and supervised by me
114. Study conceived, designed and supervised by me
115. Performed statistical analyses and data interpretation
116. Performed statistical analyses and data interpretation
117. Contributed molecular data
118. Contributed to data interpretation
119. Study conceived, designed and supervised by me
120. Study conceived, designed and supervised by me
121. Contributed to ascertainment of breast cancer families on which study based
122. Study conceived, designed and supervised by me
123. Study conceived, designed and supervised by me
124. Multi-centre study to which we contributed mutation and clinical data
125. Statistical analysis and data interpretation
126. Study conceived, designed and supervised by me
127. Study conceived, designed and supervised by me
128. Study conceived, designed and supervised by me
129. Study conceived, designed and undertaken by me and others
130. We contributed molecular data on breast tissue
131. Study conceived, designed and supervised by me
132. Study conceived, designed and supervised by me
133. Study conceived, designed and supervised by me
134. Study jointly conceived, designed and supervised with Howdle. I performed statistical analyses.
Study conceived, designed and supervised by me
Study jointly conceived, designed and supervised with Foulkes
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
We generated genotype data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
A multi-centre study to which we contributed clinico-pathological data
A multi-centre study to which we contributed clinico-pathological and mutation data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Contributed genotype data
Contributed genotype data
Study conceived, designed and supervised by me
Contributed to statistical analysis and interpretation of data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Contributed to statistical analysis and interpretation of data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Generated genotype data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Contributed mutation data and to interpretation of data
Study conceived, designed and supervised by me
Contributed to study design and data interpretation
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
A multi-centre study to which we contributed genotype data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
My research group generated genotype data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
My research group contributed mutation data
My research group genotype data, I performed linkage analysis
Study conceived, designed and supervised by me
Contributed to study design and execution
Study conceived, designed and supervised by me
My research group genotype data, I jointed directed research and interpretation of data
Study jointly conceived, designed and supervised
We performed linkage analysis of families identifying the disease locus
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Contributed to statistical analyses
Analysis of array data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
A multi-centre study to which we contributed mutation data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study jointly conceived designed and supervised with Tomlinson
I performed statistical analysis and contributed to data interpretation
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
My research group contributed to SNP array design and statistical analyses
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
A multi-centre study to which we contributed genotype data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
A multi-centre study to which we contributed to its development
Study conceived, designed and supervised by me
Contribution by virtue of performing analysis of data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study jointly conceived, designed and supervised with Tomlinson
Study conceived, designed and supervised by me
Multi-centre study to which we contributed clinicopathological data on families
Study jointly conceived designed and supervised with Tomlinson
Study jointly conceived, designed and supervised
Study jointly conceived, designed and supervised
Statistical analysis, data interpretation
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me

Study jointly conceived designed and supervised with Tomlinson

A multi-centre study to which we contributed genotype data

We performed statistical analysis and provided case samples for genotyping.

Study conceived, designed and supervised by me

Supervised statistical analysis

Study conceived, designed and supervised by me

Study jointly conceived, designed and supervised with Tomlinson

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

A multi-centre study to which we contributed genotype data

Study conceived, designed and supervised with Tomlinson

A multi-centre study to which we contributed genotype data

Study conceived, designed and supervised by me

A multi-centre study to which we contributed genotype data

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study jointly conceived designed and supervised with Tomlinson

Study conceived, designed and supervised by me

A study to which we contributed genotype data

We contributed genotype data

We contributed genotype data and statistical analysis of the association signal at 8q24 for CRC.

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

Study conceived, designed and supervised by me

We contributed linkage data to the meta-analysis

A multi-centre study to which we contributed genotype data

Study jointly conceived designed and supervised with Tomlinson

We contributed genotype data

We performed linkage analysis using SNP arrays
We performed mutation analysis of CRC tumours
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
A multi-centre study to which we contributed genotype data
A study to which we contributed genotyping information
Study jointly conceived designed and supervised with Tomlinson
Study jointly conceived, designed and supervised with Hemminki
Study jointly conceived, designed and supervised
We contributed patient samples of known mutation status
We contributed to genotyping of samples
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
We performed linkage analysis of the families
Study conceived, designed and supervised by me
We generated SNP genotypes on the cases
Study conceived, designed and supervised by me
We performed CGH analysis of the 10 patients
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
We provided breast cancer samples
We generated SNP genotypes on which study based
Contribute patient data to collaborative study
We provided patient samples to the international effort
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
We performed statistical analyses
Study conceived, designed and supervised by me
We generated and supplied GWAS SNP data on cases and control for the pooled analysis
We generated and supplied SNP data to the meta-analysis
We generated and supplied SNP data to the meta-analysis
A multi-centre study to which we contributed genotype data and performed statistical analyses
Study conceived, designed and supervised by me
Study jointly conceived, designed and supervised
A multi-centre study to which we contributed genotype data
A multi-centre study to which we contributed genotype data
Study conceived, designed and supervised by me
Study conceived, designed and supervised by me
Multicentre study jointly conceived, designed and supervised with Tomlinson and Dunlop
Assistance with statistical analyses
Study jointly conceived, designed and supervised
A multi-centre study to which we contributed genotype data
Study conceived, designed and supervised by me
Formulation of statistical analysis
We contributed to study design and genotyping
We contributed samples and mutation data to the study
328. A multi-centre study to which we contributed genotype data and performed statistical analyses
329. Contribution of samples from families and assistance with statistical analyses
330. A multi-centre study to which we contributed genotype data
331. A multi-centre study to which we contributed genotype data
332. A multi-centre study to which we contributed data from statistical analysis of association signals
333. Study conceived, designed and supervised by me
334. Study conceived, designed and supervised by me
335. A multi-centre study to which we contributed genotype data
336. A multi-centre study we jointly led
337. Study conceived, designed and supervised by me
338. Study jointly conceived, designed and supervised
339. Study conceived, designed and supervised by me
340. We generated mutation data on Hodgkin patient cases
341. Study conceived, designed and supervised by me
342. Study conceived, designed and supervised by me
343. A multi-centre study to which we contributed genotype data
344. Study jointly conceived, designed and supervised with Tomlinson
345. Study conceived, designed and supervised by me
346. A multi-centre study to which we contributed genotype data
347. Study conceived, designed and supervised by me
348. A multi-centre study to which we contributed genotype data
349. Study conceived, designed and supervised by me
350. We contributed to study design and provided assistance with genotyping and statistical analyses
351. Study conceived, designed and supervised by me
352. Study conceived, designed and supervised by me
353. Study conceived, designed and supervised by me
354. A multi-centre study to which we contributed genotype data for linkage analysis
355. Study jointly conceived, designed and supervised with Morgan
356. A multi-centre study jointly led with Tomlinson and Dunlop
357. Study jointly conceived, designed and supervised with Swerdlow
358. Study conceived, designed and supervised by me
359. A multi-centre study jointly led by us
360. Study jointly conceived, designed and supervised with Dunlop and Tomlinson
361. A multi-centre study to which we contributed genotype data
362. Sole author
363. Contributor of clinical data
364. We provided assistance with statistical analysis and supplied clinicopathological data
365. Study conceived, designed and supervised by me
366. A multi-centre study to which we contributed genotype data
367. Study jointly conceived, designed and supervised with Dunlop and Tomlinson
368. Study jointly conceived, designed and supervised with Slager and Skibola
369. A multi-centre study to which we contributed genotype data
370. Contributor to data analysis
371. Contributor to data analysis
372. Study conceived, designed and supervised by me.
A multi-centre study to which we contributed genotype data
A multi-centre study to which I provided statistical assistance
Study jointly conceived, designed and supervised by me
A multi-centre study to which we contributed genotype data
Study jointly conceived, designed and supervised with Amos and Landi
Study conceived, designed and supervised by me
A multi-centre study to which we contributed genotype data
Study conceived, designed and supervised with Amos and Landi
Study conceived, designed and supervised by me
A multi-centre study to which we contributed genotype data

BOOK CHAPTERS

B1. One of seven contributors
B2. Co-written equally by the 4 authors
B3. Co-written equally by the 4 authors
B4. Co-written equally by the 2 authors
B5. Co-written equally by the 2 authors
B6. Sole author
B7. Co-written equally by the 2 authors
B8. Co-written equally by the 2 authors
B9. Co-written equally by the 2 authors
B10. Co-written equally by the 2 authors
B11. Co-written equally by the 2 authors
B12. Co-written equally by the 3 authors
B13. Co-written equally by the 2 authors
B14. Co-written equally by the 2 authors
B15. Co-written equally by the 2 authors
B16. Co-written equally by the 3 authors
B17. Co-written equally by the 3 authors
B18. Co-written equally by the 2 authors
B19. One of the five contributors
B20. Co-written equally by the 2 authors
B21. Co-written equally by the 3 authors
B22. Co-written equally by the 3 authors
B23. Co-written equally by the 2 authors
B24. One of the five contributors
B25. One of the four contributors
B26. Co-written equally by the 3 authors
B27. Co-written equally by the 3 authors
B28. Co-written equally by the 2 authors
B29. Co-written equally by the 2 authors
B30. Co-written equally by the 2 authors
B31. Co-written equally by the 3 authors
B32. Co-written equally by the 2 authors
B33. Co-written equally by the 2 authors
B34. Co-written equally by the 3 authors

REVIEWS
R1. Co-written equally by the 2 authors
R2. Co-written equally by the 2 authors
R3. Co-written equally by the 2 authors
R4. Co-written equally by the 2 authors
R5. Co-written equally by the 2 authors
R6. Co-written equally by the 2 authors
R7. Co-written equally by the 2 authors
R8. Co-written equally by the 2 authors
R9. Co-written equally by the 2 authors
R10. Sole author
R11. Co-written equally by the 2 authors
R12. Co-written equally by the 2 authors
R13. Sole author
R14. Co-written equally by the 2 authors
R15. Co-written equally by the 3 authors
R16. Co-written equally by the 4 authors
R17. Co-written equally by the 2 authors
R18. Sole author
R19. Co-written equally by the 2 authors
R20. Co-written equally by the 2 authors
R21. Co-written equally by the 3 authors
R22. Co-written equally by the 5 authors
R23. Co-written equally by the 2 authors
R24. Co-written equally by the 2 authors
R25. Co-written equally by the 3 authors
R26. Co-written equally by the 2 authors
R27. One of the eight contributors
R28. Co-written equally by the 3 authors
R29. Co-written equally by the 2 authors
R30. Co-written equally by the 2 authors
R31. Co-written equally by the 2 authors
R32. Co-written equally by the 2 authors
R33. Sole author
R34. Co-written equally by the 4 authors
R35. One of the six contributors
R36. Sole author
R37. Co-written equally by the 2 authors
R38. Co-written equally by the 2 authors
R39. Co-written equally by the 2 authors
R40. Co-written equally by the 3 authors
R41. Co-written equally by the 3 authors

LETTERS

L1. Co-written equally by the 2 authors
L2. Co-written equally by the 3 authors
L3. One of the eight contributors
L4. Co-written equally by the 2 authors
L5. Co-written equally by the 3 authors
L6. Co-written equally by the 2 authors
List of the publications


