MINIREVIEW

Walter F. Bodmer

Cancer genetics: colorectal cancer as a model

Abstract Cancer is essentially a somatic evolutionary process and is, therefore, effectively defined by the genetic and epigenetic changes underlying this process. An understanding of the function of these changes is fundamental to devising new approaches to prevention and treatment. Colorectal cancer (CRC), apart from its obvious importance as one of the most frequent cancers, provides an excellent model for such studies because of the availability of precursor adenoma lesions and the existence of several clear-cut familial inherited susceptibilities. These include familial adenomatous polyposis (FAP), which led to the identification of the APC gene and the importance of the Wnt pathway, and hereditary non-polyposis CRC (HNPCC), which identified the role of the mismatch repair genes in colorectal and other cancers. The presently known range of genetic and epigenetic changes in CRCs and adenomas is reviewed in this paper and the evidence against a requirement for genomic instability presented, together with a discussion of patterns of gene methylation, including especially our work on the homeobox gene, CDX1. Clearly, familial cancers, such as FAP and HNPCC, cannot account for more than perhaps 5% of the incidence of CRC. There is, however, evidence that approximately a further 25–30% have some inherited susceptibility. Based on the association of APC missense variants with multiple adenomas, we proposed that much of this may be due to the cumulative effects of low frequency, low penetrance variants, and the "rare variant hypothesis". The evidence for this from our work on multiple adenoma cases, and certain other examples, is discussed.

W.F. Bodmer
CR-UK Cancer and Immunogenetics Laboratory,
Weatherall Institute for Molecular Medicine,
Oxford University, Oxford, OX3 9DS, UK
E-mail: walter.bodmer@hertford.ox.ac.uk
Tel.: +44-1865-222356
Fax: +44-1865-222431

Introduction

Colorectal cancer (CRC) provides a good model both for the study of disease susceptibility and for the somatic evolution of epithelial cancers. This is because there are well-defined inherited susceptibilities to CRC and because the different stages of CRC are more readily accessible than for most other carcinomas.

The best known Mendelian, clearly familial, syndrome is familial adenomatous polyposis (FAP). Affected individuals have from a few hundred to more than a thousand pre-cancerous growths, namely polyps or adenomas. CRC is inevitable by the age of 20 or 30. The first clear FAP families were described in 1925 by the surgeon Lockhart-Mummery who recognised the pattern of Mendelian inheritance. FAP is a classical, rare, dominantly inherited disease maintained in the population at a frequency of about 1/8,000 by mutation selection balance. Until recently, it was clearly severely disadvantageous so that most affected individuals would have died before reproductive age.

Loss of heterozygosity in sporadic CRC provided the initial evidence that the gene responsible for FAP was involved in somatic as well as germ line changes in cancers, as predicted by Knudson in 1971.

Familial adenomatous polyposis was one of the first relatively common Mendelian diseases whose gene was identified by positional cloning, based on the identification by Nakamura of a sufficiently small deletion. The resulting APC gene turns out to be a most important gene for all CRC. Thus, about 80% of all CRCs have their first mutation in this gene, which is therefore a key initiator of carcinogenesis. Many *APC* mutations have now been identified, both in the germ line in FAP patients, and somatically in sporadic CRCs. There are two relatively common germ line mutations at two different 5 base pair repeat positions where the mutation is up to 1,000 times that at ordinary single base pair changes. CpG positions, often with the C methylated, have about 40 times that rate. This information can be used to cover

the commoner mutations for FAP family screening of the *APC* gene.

Nakamura and his colleagues identified in somatic *APC* mutations a "mutation cluster region" (MCR) of the gene, which accounted for about 60% of all somatic mutations. The evidence strongly suggests that this is the result of functional selection for mutations in the MCR and not a mutation rate effect. In particular, it has been shown both for sporadic tumours and tumours in FAP patients that, when the first mutation is in the MCR, then the second event leading to loss of the normal functioning version of the *APC* gene is due to loss of heterozygosity, mainly as a result of mitotic recombination. On the other hand, when the first mutation is not in the MCR, then the second event is a further mutation in the MCR. This implies that there is a selective advantage for mutations in the MCR, effectively a dominant negative effect probably due to major disruption of the APC protein's interactions with other proteins (for a review of FAP and the *APC* gene see Fearnhead et al. 2001)

The second major inherited CRC susceptibility is hereditary non-polyposis CRC, or HNPCC. Families with HNPCC have mutations in mismatch repair genes, mainly *hMLH1* and *hMSH2*. As a result most tumours in HNPCC patients have microsatellite instability (MSI) recognised by novel mutations in microsatellite markers in the patient's tumours, but not in their germ line DNA. MSI also occurs in about 15% of sporadic tumours, mainly by hMLH1 methylation silencing rather than by mutation, as in the familial cancers (see for example Fearnhead et al. 2002).

The adenoma to carcinoma sequence

The definition of an adenoma to carcinoma sequence in CRC was first made by the pathologist, Basil Morson, working in London at St Mark's Hospital where FAP had first been clearly identified. Once some of the genes mutated in CRC were identified, it became clear that these could, to some extent, be fitted into this adenoma to carcinoma sequence (Fig. 1). Thus, *APC* mutations occurred in early adenomas, while, for example, mutations in the *TP53* gene occurred later in the sequence, perhaps mostly at the late adenoma or early carcinoma stage, but well before metastasis.

Mutated genes in sporadic CRC include those involved in the Wnt pathway, namely APC, beta catenin and E-cadherin, genes involved in apoptosis including p53 and the mismatch repair genes, *hMLH1*, *hMSH2* and *hMSH6*, cell cycle check point genes such as p14, p16, genes for growth factors, their receptors and signalling (TGFbetaIIR, SMAD4, Kras) and genes involved in immune attack resistance such as beta2m and HLA Class I.

Using CRC-derived cell lines it is sometimes possible to identify patterns of association between mutations in certain sets of genes that point to different evolutionary pathways. For example, MSI tumours mostly have mutations in TGFβIIR, while those without mismatch repair deficiency frequently have mutations in SMAD4 and/or 18qLOH, but not mutations in TGFβIIR. This implies that there are contrasting evolutionary pathways in MSI and non-MSI tumours (Woodford-Richens et al. 2001).

Colorectal cancer-derived cell lines can enable a more precise characterisation of the mutational spectrum for a gene than studies on primary tumours. This is illustrated by a recent comprehensive analysis of the *TP53* gene and its protein status in a panel of 56 CRC cell lines. This analysis was based on a combination of mutation screening of all the exons of the p53 gene, cDNA sequencing, and the assessment of the function of the p53 protein by assaying the induced expression of phosphorylated p53 and p21 after exposing cells to γ rays. Thirteen of the 56 cell lines had functional p53, 21 lines had missense mutations (one of which made no detectable protein), 4 lines produced no p53 transcripts, and the remaining 18 lines carried truncating*TP53* mutations. The results showed a relatively high frequency of *TP53* mutations (76.8%) in the cell lines with almost half of the mutations being truncating mutations, which is a rather higher frequency of such mutations
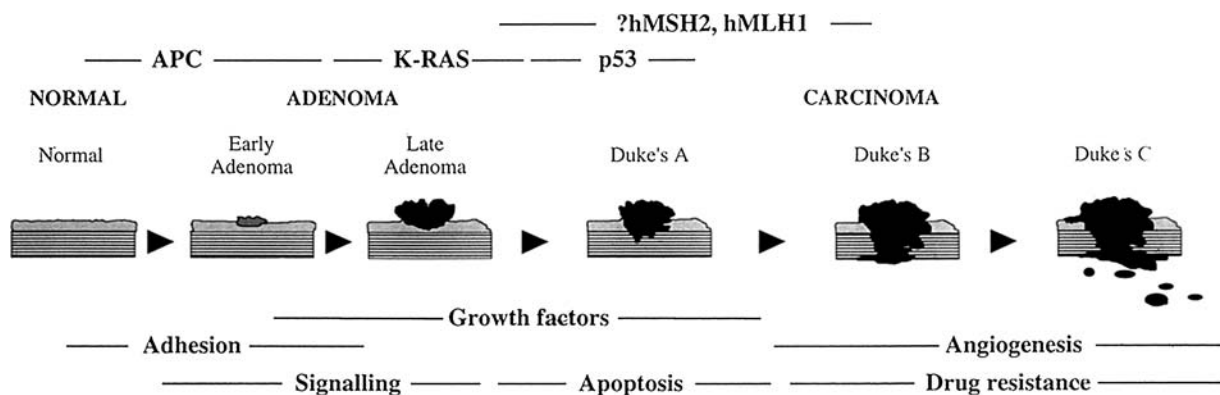

Fig. 1 Adenoma to carcinoma sequence

than usually reported. Of the 18 cell lines with truncating mutations, 12 had detectable truncated protein based on western blot analysis, while no protein was detected in the remaining 6 cell lines. The results raise the question of the extent to which truncating mutations may have dominant negative effects, even when no truncated protein can be detected by standard methods (Li and Bodmer 2006).

There is no requirement for genomic instability in cancers

It has often been argued that there is a requirement for genomic instability in the somatic evolution of sporadic cancers. The arguments are based on the idea that there are too many genetic changes for normal mutation rates to be sufficient for carcinogenesis. They appear to be supported by the observation that tumours often do acquire higher mutation rates and chromosomal instability.

However, there are very strong arguments against such a requirement for genomic instability in cancers. Thus, it has been shown that mismatch repair mutations do not occur before *APC* mutations (Homfray et al. 1998), whereas it is exactly at this early stage of tumour development that one would expect a need for an increased mutation rate according to the instability hypothesis. Furthermore, only a limited range of repair genes is found mutated in cancers and not all cancers are chromosomally unstable. In addition, it has been shown that chromosomal instability is not an early event, as again would be expected if instability was a requirement for carcinogenesis (Sieber et al. 2002).

Theoretical models show that selection overrides mutation rate as an evolutionary driving force in cancers, as it does in organismal evolution. Ignoring the power of Darwinian natural selection is a key feature of those who argue for the need for genomic instability in the development of cancers.

Finally, most so-called tumour suppressor genes are not truly recessive at the level of the tumour, as is evident for both *APC* and *TP53*. In both these cases there is clear evidence for a selective advantage of the first mutation, even though selection nearly always leads subsequently to complete loss of the gene's normal function.

Expression control changes by methylation in CRC: CDX1 as an example

Relatively stable epigenetic changes in gene expression, such as occur during cellular differentiation, have been recognised for many years as potentially able to play a comparable role to genetic mutations in the somatic evolution of cancers. More recently, this has been clearly shown in many cases as a result of the development of techniques for identifying expression control through gene promoter methylation, which generally results in switching off gene expression. The effect of methylation is then comparable to inactivating gene mutations, though it cannot give rise to dominant negative effects, since it only simply reduces the level of production of a gene product, generally by half. In that case, an advantage for an initial change in just one copy of a gene must come from haplo-insufficiency. In contrast to gene mutation, there may, however, be mechanisms that in some cases methylate more or less synchronously the promoters of both versions of a gene in the diploid state. Examples of genes in which promoter methylation has been shown to be a mechanism probably selected during tumourigenesis include those for the epithelial attachment protein, E-Cadherin, the mismatch repair gene, *hMLH1*, and the cell cycle control proteins p14 and p16. The methylation of the homeobox gene *CDX1*, which controls key aspects of the differentiation of the colonic epithelium, will be discussed as an example.

The key technology used is bisulfite DNA sequencing, which recognises which C residues in the DNA are methylated (Frommer et al. 1992). Areas of interest in the gene are amplified, bisulfite modified, cloned and at least ten clones from each source of DNA sequenced to establish the pattern of methylation. An example of the results obtained for three CRC-derived cell lines is illustrated in Fig. 2. This shows a very clear correlation between promoter methylation and lack of CDX1 mRNA expression (in the cell line C10). However, the results are not always so clear-cut, and the evidence suggests that it is just certain key CpG positions in the promoter, often at transcription factor binding sites, that are critical for the control of gene expression. There are also situations where the frequency of methylation and the level of gene expression seem to vary with the culture conditions. This may, for example, represent changes in the proportion of partially differentiating cells in the culture. The normal colon uniformly expresses CDX1 and shows no methylation of the gene's promoter.

There are two categories of tumour changes, one in which there is full methylation of the CDX1 promoter and no CDX1 expression, and the second in which there is partial methylation of the CDX1 promoter and variable, generally lower, CDX1 expression. In both cases it is presumed that reduced CDX1 expression is selected perhaps because of reduced constraints on epithelial cellular differentiation (Wong et al. 2004).

Inherited susceptibility to CRC: the rare variant hypothesis

A scheme for types of CRC-inherited susceptibility is shown in Fig. 3. Chance and the environment probably account for at least 70% of all sporadic cases, while clear-cut Mendelian inherited syndromes account for no more than about 5%. The remainder, perhaps about 25%, represents a "multifactorial" contribution that is
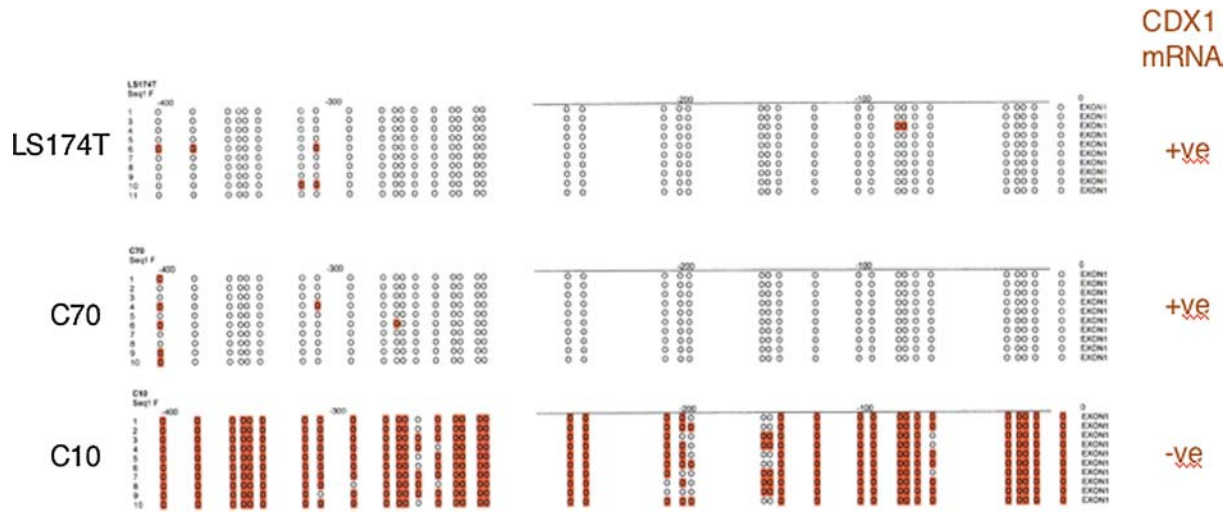
Fig. 2 CDX1 promoter methylation and mRNA expression for three colorectal cancer cell lines. Each *small circle* represents the position of a CpG site in the promoter region. *Filled circles* indicate methylated Cs. The results from ten clones are shown for each cell line

Fig. 3 Scheme for inherited susceptibility to colorectal cancer
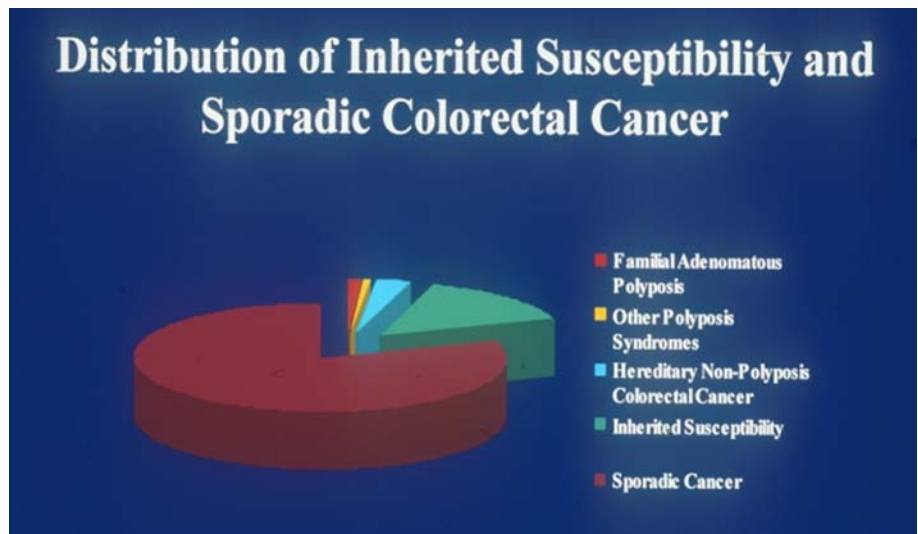
not associated with clear-cut Mendelian families. This is where there is the biggest gap in our knowledge.

Important clues to understanding the nature of the multifactorial inherited component of CRC, and in general of any chronic disease, have come from the discovery of APC missense variants that predispose to CRC, but do not in general give rise to familial clustering. Thus, the APC missense variant I1307K was found by Laken et al. (1997) to be at a significantly higher frequency in Ashkenazi Jewish colorectal "families" than in a control population. With the aim of confirming and extending these results, Frayling et al. 1998 looked for missense variants in the MCR of the APC gene in a group of colorectal adenoma patients and a smaller set of early onset CRC cases that were not obviously familial, and certainly not either FAP or HNPCC. They found a significantly increased frequency of the I1307K variant in a small number of Ashkenazi

Jewish sporadic adenomas and early onset cancer cases compared with controls, in support of Laken et al.'s familial CRC results. However, they found, in addition, another missense variant, E1317Q, which was not present in the Jewish patients and which also showed a significantly increased presence in both the adenoma and early onset cancer cases. These data were subsequently confirmed and extended to a larger sample of adenoma cases (Lamlum et al. 2000).

The results of these studies raise the question of the nature of such "rare variants" with dominant effects. These variants are not obviously polymorphic, but are too common to be maintained by mutation–selection balance. Such dominantly acting rare variants, associated with increased disease risk, but at a low penetrance level, can simply increase by chance because they will be subject to hardly any adverse selection and so initially will behave like neutral mutations, which have a small

chance of increasing to a significant frequency. These variants may, in aggregate, make a substantial contribution to inherited susceptibility, in this case to adenomas and early onset CRC.

These ideas led to the general proposal of the ''rare variant hypothesis'', first made in outline by Frayling et al. 1998 and subsequently discussed in more detail by Bodmer (1999), as generally applicable to many cases of multifactorial inherited susceptibility. Thus, much of the inherited susceptibility to human chronic diseases in general may be due to the summation of the effects of a series of low frequency dominantly acting variants of a variety of different genes, each conferring a moderate, but readily detectable, increase in relative risk. There are now many such examples with relative risks greater than approximately 2. The rare variants will mostly be population-specific because of founder effects due to genetic drift.

The alternative hypothesis is that there are common polymorphic alleles of a limited number of genes, which can be detected by association studies with polymorphic single nucleotide polymorphisms (SNPs) due to linkage disequilibrium (LD). This is the whole basis for SNP association analysis, whose origin is in the widely studied HLA (human major histocompatibility system) and disease associations (see e.g. Tomlinson and Bodmer 1995). There are, however, few examples of success with this approach and most have given rise to relative risks of less than 1.5. The exceptions seem to be due to variation associated with prior selection, as for HLA, and often perhaps in early human populations.

Simple calculations show that, even with a 50% penetrance, most of those affected due to a dominant variant will occur in families as sporadic cases. This is because the chance of two members of a family being affected is small, as it requires both the presence of the gene and its expression in terms, for example, of causing adenomas. One cannot therefore rely on studying families to find such rare variants. The only secure way to find them is by DNA sequencing candidate genes in selected patient groups.

In my laboratory we have extended our studies on multiple colorectal adenomas and used the well-known presumed susceptibility to such adenomas as a model to test the rare variant hypothesis (Fearnhead et al. 2004). Our patient group was a series of 124 individuals with 3–100 adenomas that did not include any obvious familial cases. Candidate genes were chosen either because severe mutations, as in APC and HNPCC, cause clear-cut familial disease, or because mutations or epigenetic somatic changes are known to play a part in the progression to sporadic CRC. The choice of candidates depends on a good knowledge of both the CRC pathway and of inherited susceptibilities to CRC. Each candidate gene is then scanned for variants in its DNA sequence in all the adenoma patients. When a variant is found, an assay is set up for it in order to test for its presence in a control population made up of random individuals who are, however, not necessarily known not to have polyps. Individual variants are often not significantly more frequent in cases than controls because the numbers are relatively small. There will undoubtedly be some individuals in the controls with multiple adenomas, and this dilutes the level of association and thus the chance of finding a significant effect. It is, however, striking, as shown in Table 1, that all the variants found are putatively functional. Many of those, for example, in the mismatch repair genes, *hMLH1* and *hMSH2* have actually been shown to be so in yeast model systems. In contrast, the commoner *hMLH1* polymorphisms are, for example, all putatively neutral occurring either as synonymous third base pair changes or in introns not near to splice junctions.

The combined statistical analysis demonstrates a very significant overall excess of the variants in the patients compared with the controls. Each variant can be assessed as to its likely functional effect and essentially all satisfy the criteria for this. Thus, they occur, for example, in conserved regions, in known functionally relevant regions, or involve charge changes.

There are now also striking published examples of rare variants for other cancers, for example, for *BRCA2* in prostate cancer, and for *Chk2* in breast cancer (for a review see for example Fearnhead et al. 2005).

This idea of rare variants underlying the basis of multifactorial susceptibility may apply to many common chronic diseases such as heart and mental diseases including Alzheimer's, diabetes and auto-immune dis-

Table 1 Summary of variants found in germline DNA from 124 patients with multiple adenomas and 483 random controls

Gene	Variants	Patients (%)	Controls (%)	Odds ratio	P
APC	*E1317Q*	3/124 (2.4)	6/480 (1.3)	2.0	0.400
CTNNB1	N287S	1/124 (0.8)	3/483 (0.6)	1.3	1.000
AXIN1	P312T, R398H, L445M, D545E, G700S, R891Q	18/124 (14.5)	37/479.3[a] (7.7)	2.0[c]	0.012
hMLH1	G22A, K618A	5/124 (4.0)	11/482.5 (2.3)	2.0[c]	0.175
hMSH2	H46Q, E808X, ex4SDS	3/124 (2.4)	0/479.3 (0.0)	11.7[c]	0.001
	Combined	30/124 (24.9)	55/479.8[a, b] (11.5)	2.2[c]	0.0001

SDS splice donor site variant
[a] Average number of controls typed for all variants
[b] Two individuals each had two variants
[c] Modified Mantel–Haenszel odds ratios

eases. The variants simply increase by chance until counter-selection prevents further increase.

In summary, the key to identifying candidate genes, and so rare dominantly acting variants, is to look in genes where there is an a priori basis for expecting an effect or where mutations with a severe effect on protein function have been shown to have analogous, but much more severe clinical effects, as in the case of FAP and the *APC* gene. The variants will be most likely to lead to amino acid changes that affect protein–protein interactions, and so can have dominant or dominant negative effects. Variants in promoter regions may sometimes have dominant effects on levels of gene expression and so may also be important in this context. The selective disadvantage of such variants will be minimal, especially in relation to the common chronic diseases of the present, which were hardly of significance until the quite recent past. Such variants will, thus, from time to time, drift up in frequency by chance, and may then be seen as "founder" variants such as *CTNNB1 N287S* or *hMLH1 K618A* in the adenoma study.

Collectively, it may be worth screening for such rare variants in CRC, since colonoscopy should then give a higher than usual chance of finding an adenoma, which, when removed, eliminates the increased risk due to the presence of the adenoma(s). This is a comparatively non-invasive procedure with minimal morbidity. The DNA-based technology for screening for known genetic variation is comparatively straightforward and is bound to keep going down in cost. However, even if a variant is too rare to form a basis for screening, it may nevertheless provide valuable information about disease aetiology.

References

Bodmer WF (1999) Familial adenomatous polyposis (FAP) and its gene, APC. Cytogenet Cell Genet 86:99–104

Fearnhead NS et al (2001) The ABC of APC. Hum Mol Genet 10(7):721–733

Fearnhead NS, Wilding JL, Bodmer WF (2002) Genetics of colorectal cancer: hereditary aspects and overview of colorectal tumorigenesis. Br Med Bull 64:27–43

Fearnhead NS, Wilding JL, Winney B, Tonks S, Bartlett S, Bicknell DC, Tomlinson IPM, Mortensen NJ, Bodmer WF (2004) Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. Proc Natl Acad Sci USA 101:15992–15997

Fearnhead NS, Winney B, Bodmer WF (2005) Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model. Cell Cycle 4:521–525

Frayling IM, Beck NE, Ilyas M, Dove-Edwin I, Goodman P, Pack K, Bell JA, Williams CB, Hodgson SV, Thomas HJW, Talbot IC, Bodmer WF, Tomlinson IPM (1998) The APC variants I1307K and E1317Q are associated with colorectal tumors, but not always with a family history. Proc Natl Acad Sci USA 95:10722–10727

Frommer M, McDonald LE, Millar DS, Collis CM, Watt F, Grigg GW, Molloy PL, Paul CL (1992) A genomic sequencing protocol that yields a positive display of 5-methylcytosine residues in individual DNA strands. Proc Natl Acad Sci USA 89:1827–1831

Homfray TF et al? (1998) Defects in mismatch repair occur after APC mutations in the pathogenesis of sporadic colorectal tumours. Hum Mutat 11(2):114–120

Laken SJ, Petersen GM, Gruber SB, Oddoux C, Ostrer H, Giardiello FM, Hamilton SR, Hampel H, Markowitz A, Klimstra D, Jhanwar S, Winawer S, Offit K, Luce MC, Kinzler KW, Vogelstein B (1997) Familial colorectal cancer in Ashkenazim due to a hypermutable tract in APC. Nat Genet 17:79–83

Lamlum H, Al Tassan N, Jaeger E et al (2000) Germline APC variants in patients with multiple colorectal adenomas, with evidence for the particular importance of E1317Q. Hum Mol Genet 9(15):2215–2221

Liu Y, Bodmer WF (2006) Analysis of P53 mutations and their expression in 56 colorectal cancer cell lines. Proc Natl Acad Sci USA 103:976–981

Sieber OM, Heinimann K, Gorman P, Lamlum H, Crabtree M, Simpson CA, Davies D, Neale K, Hodgson SV, Roylance RR, Phillips RKS, Bodmer WF, Tomlinson IP (2002) Analysis of chromosomal instability in human colorectal adenomas with two mutational hits at APC. Proc Natl Acad Sci USA 99(26):16910–16915

Tomlinson IP, Bodmer WF (1995) The HLA system and the analysis of multifactorial genetic disease. Trends Genet 11:493-498

Wong NA, Britton MP, Choi GS, Stanton TK, Bicknell DC, Wilding JL, Bodmer WF (2004) Loss of CDX1 expression in colorectal carcinoma: promoter methylation, mutation, and loss of heterozygosity analyses of 37 cell lines. Proc Natl Acad Sci USA 101(2):574–579

Woodford-Richens KL, Rowan AJ, Gorman P, Halford S, Bicknell DC, Wasan HS, Roylance RR, Bodmer WF, Tomlinson IP (2001) SMAD4 mutations in colorectal cancer probably occur before chromosomal instability, but after divergence of the microsatellite instability pathway. Proc Natl Acad Sci USA 98(17):9719–9723