NIH Public Access
Author Manuscript
Surg Oncol Clin N Am. Author manuscript; available in PMC 2010 October 1.

Published in final edited form as:
Surg Oncol Clin N Am. 2009 October ; 18(4): 663–668. doi:10.1016/j.soc.2009.07.004.

Genome-Wide Association Studies and Colorectal cancer

Loïc Le Marchand, M.D., Ph.D.
Professor, Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, Hawaii. This work was supported by grants 1R01CA126895 and 1U01HG004802 from the National Institutes of Health. Key Words: Colorectal cancer, Genetics, GWAS

Abstract

Genome-wide association studies (GWAS) provide a powerful new approach to identify common, low-penetrance susceptibility loci without prior knowledge of biological function. Results from three GWAS conducted in populations of European ancestry are available for colorectal cancer (CRC). These studies have identified eleven disease loci which, for the majority, were not previously suspected to be related to CRC. The proportions of the familial and population risks explained by these loci are small and they currently are not useful for risk prediction. However, the power of these studies was low indicating that a number of other loci may be identified in new on-going GWAS, and in pooled analyses. Thus, the risk prediction ability of susceptibility markers identified in GWAS for CRC may improve as more variants are discovered. This may in turn have important implications for targeting high risk individuals for colonoscopy screening.

Introduction

Colorectal cancer (CRC) is known to aggregate in families, with the disease being two-to-three times more common among the first degree-relatives of cases than in those of population controls. The contribution of inherited factors (mainly genetic) to the etiology of the disease has been estimated in twin studies to be 35% (1). However, for the most part, the underlying susceptibility genes for CRC remain unknown. In recent decades, linkage studies used collections of multi-case families to identify a number of rare mutations in highly penetrant genes that cause well characterized Mendelian syndromes (e.g., HNPCC, FAP, juvenile polyposis, Peutz Jeghers syndrome) (2). However, these mutations explain 2–6% of CRCs, and only a small fraction of the familial risk. Thus, it is likely that additional susceptibility genes exist for CRC.

In recent years, linkage studies have failed to discover additional high-penetrance genes, suggesting that multiple low-penetrance alleles may explain the remaining genetic risk for CRC. Indeed, association studies, in which the frequencies of genetic variants are directly compared between large series of patients and unrelated controls , are now thought to be more appropriate than linkage studies for the identification of susceptibility loci for complex diseases, including CRC (3).

Address for correspondence: Loïc Le Marchand, M.D., Ph.D., Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, 1236 Lauhala Street, Suite 407, Honolulu, HI 96813, (808) 586-2988, (808) 586-2982 (fax), loic@crch.hawaii.edu (email).

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Human Genetic Variation and the Study of Complex Diseases

It is estimated that there are ~10 million single-nucleotide polymorphisms (SNPs) in the human genome, half of which with a minor allele frequency (MAF) over 10% (4). These genetic variants and other types of polymorphisms (insertion/deletion, copy number variations) are expected to explain approximately 90% of human heterozygosity, including susceptibility to disease (4). Variants that were deleterious during evolution (such as mutations that cause early-onset diseases) are typically rare, due to natural selection. Conversely, disease variants that act after reproduction, or that are pleiotropic in effect, may have been neutral or subject to balancing selection (e.g., sickle cell anemia and malaria). In such cases, most of the genetic variation underlying disease risk may be common.

Detailed studies of the variation in the human genome across individuals found sizeable regions, or “linkage disequilibrium (LD) blocks”, over which little evidence for past recombination was observed, and within which more than 90% of all chromosomes matched to only one of a few common haplotypes (5). These studies showed that nearly all of the common diversity at a given locus could be captured by genotyping a small subset of common markers.

Over the past seven years, international efforts have resulted in a public reference human genome diversity database, a Haplotype Map of the Human Genome (HapMap) which has identified and validated over 3 million SNPs (6). These resources, as well as the development of high-throughput microarray platforms for the simultaneous genotyping of hundreds of thousands of SNPs now allow the testing of a high proportion of all common SNPs (with frequency >5%) for association with disease in studies called “genome-wide association studies” (GWAS). These studies allow the scanning of the entire genome for association with disease without prior knowledge of biological function and, thus, have the potential to reveal unsuspected regions and novel biological mechanisms. However, theses studies require large sample sizes to account for the inflated Type-I error resulting from the very high number of case-control comparisons and to detect effect sizes that are expected to be small.

Published GWAS of Colorectal Cancer

Over the last two years, results from GWAS have been published for CRC. These studies have all been case-controls studies conducted in populations of European ancestry and used multistage designs (7–13). Table 1 summarizes the published findings from these studies as of January 2009. The risks conferred by each risk allele have uniformly been low, with odds ratios in the range of 1.1–1.3 per allele (7–13).

The first susceptibility locus for CRC identified in these studies was 8q24. This genomic region first emerged for prostate cancer, through a linkage study followed up by an association study and, independently through an admixture scan in African Americans (14,15). At least three different susceptibility loci were identified for prostate cancer in 8q24 (16). One of these loci (128.1–128.7 Mb) was found to be associated with CRC in two GWAS (7,8) and, independently, in a case-control study nested in the Multiethnic Cohort study (17). In this region, subsequently also associated with ovarian cancer, there are no known genes or annotated coding transcripts, with the exception of a pseudogene (*POU5F1P1*). However, approximately 300 kb telomeric to this region is the c-*MYC* (*MYC*) oncogene. Replication, sequencing and fine-mapping studies of this locus have identified rs6983267 as the most promising variant for functional assessment (18). This SNP lies in a sequence which is highly conserved across vertebrates and is predicted to have regulatory function (18). Although *MYC* is often amplified in colon and prostate cancers, rs6983267 has not been found to modify the expression of this gene in colon tumors and lymphoblastoid cell lines. Thus, the mechanism underlying the association of this SNP to CRC and several other common cancers remains

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

unknown. However, its relative proximity to *MYC* makes it plausible that it may disrupt one of its putative distant enhancers, the effect of which, however, may not be observable in tumors.

A locus at 9p24, also a region with no obvious candidate gene, was also found associated with CRC in the original ARCTIC report (7) and was replicated in the Colorectal Cancer Family Registry (19). However, since this association was not observed in some of the ARCTIC replication populations, this association may not exist in all populations.

A number of the subsequently reported loci fall within or close to a gene (18q21: *SMAD7*; 15q13.3: *CRAC1*, 8q23.3: *EIF3H*; 14q22.2: *BMP4*; 16q22.1: *CDH1*; and 19q13.1: *RHPN2*). SMAD7 is known to act as an intracellular antagonist of TGF signaling and perturbation of its expression had been shown to affect CRC progression (9). EIF3H regulates cell growth and viability (11). *CRAC1* had already been linked to hereditary mixed polyposis syndrome and CRC in Ashkenazi Jews (10). However, the other associated loci (10p14, 11q23.1, 18q23, 20p12.3), similarly to 8q24 and 9p24, lie in intergenic regions with no known biological relevance. Thus, a large amount of work is needed to understand the biological mechanisms underlying these associations.

Research Needs

However, before functional studies can be initiated, re-sequencing and fine-mapping efforts are needed to identity the best candidate causal variants at these newly identified eleven loci. Moreover, very little information is also available on the generalization of these associations to ethic/racial groups other than whites. The only exception is rs6983267 at 8q24, which has been shown to be consistently associated with CRC among the five ethnic/racial populations in the Multiethnic Cohort (Japanese Americans, Native Hawaiians, African Americans, Whites and Latinos) (17) and to be the best candidate variants in the region (18). Tenesa et al. (12) have also suggested that rs3802842 at 11q23 may not be associated with CRC in Japanese. Fine-mapping studies in populations with different LD structures are potentially very useful to identify the true causal variant at a particular locus, as well as novel ethnic/racial-specific risk alleles.

Only limited data are available on the epidemiological characteristics of these associations. Rs3802842 at 11q23 and rs4939827 (*SMAD7*) have been reported to be more strongly associated with rectal cancer than colon cancer (12). No differences in risk have been reported by tumor molecular subtypes for the eleven published variants, with the exception of rs4444235 (*BMP4*) for which the association was found to be significantly stronger for mismatch repair (MMR) proficient tumors than for MMR deficient tumors (13). The largest analysis conducted to date, a pooled analysis of the two UK studies, have suggested that each of the risk alleles identified so far have independent effects and that, as a group, they only explain a small proportion of cases in the population (13). However, even in this pooled analysis, power to detect associations with SNPs having a MAF <0.3 was limited. This suggests that additional, somewhat less common, susceptibility variants exist and points to the need for a pooled analysis with the ongoing North American GWAS studies and the need to conduct additional GWAS.

The potential modifying effects of the newly identified susceptibility variants also need to be investigated. It is very clear from migrant and temporal trend studies that the etiology of CRC has a very strong environmental component (20). Thus, large cohort studies, in which lifestyle risk factors for CRC were assessed before diagnosis, are being used to investigate gene-environment (GxE) interactions with the risk alleles identified in GWAS. Pooled analyses of published and existing GWAS may also provide adequate power to detect novel modifying genes in investigations of GxG interactions in the primary data (21). Finally, populations that are especially susceptible to the effect of a Western lifestyle on CRC risk, such as the Japanese, may provide a particularly suitable population for identifying GxE interactions (20,22).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Summary

GWAS provide an efficient new approach to identify common, low-penetrance susceptibility loci without prior knowledge of biological function. Results from GWAS conducted in populations of European ancestry living in the UK and Canada have been published for CRC. These studies have identified eleven well-replicated disease loci which, for the majority, were not previously suspected to be related to CRC. "Post-GWAS" studies are being initiated to: 1) characterize the epidemiology of these associations across populations, tumor molecular subtypes and clinically relevant variables; 2) explore gene-environment interactions to detect modifying effects that may explain a greater proportion of the population risk; and 3) identify the best candidate causal variants for subsequent functional studies aimed at elucidating the underlying biological mechanisms. The proportions of the familial and population risks explained by the published loci are small and they are not currently useful for risk prediction. However, the power of the published studies was low indicating that a number of other loci may be found in additional ongoing GWAS, especially as the result of pooled analyses of all the combined primary data. Thus, there is potential for the risk prediction ability of susceptibility markers identified in GWAS to improve as more variants are found. This may in turn have important implications for targeting high risk individuals for colonoscopy screening.

References

1. Lichtenstein P, Holm NV, Verkasalo, 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;343:75–85.
2. Burt RW, DiSario JA, Cannon-Albright L. Genetics of colon cancer: impact of inheritance on colon cancer risk. Annu Rev Med 1995;46:371–9. [PubMed: 7598472]
3. Risch NJ. Searching for genetic determinants in the new millennium. Nature 2000;405(6788):847–56. [PubMed: 10866211]
4. Kruglyak L, Nickerson DA. Variation is the spice of life. Nat Genet 2001;27(3):234–6. [PubMed: 11242096]
5. Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. Science 2002;296(5576):2225–9. [PubMed: 12029063]
6. Frazer KA, Ballinger DG, Cox DR, et al. International HapMap Consortium. A second generation human haplotype map of over 3.1 million SNPs. Nature 2007;449(7164):851–61. [PubMed: 17943122]
7. 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(8):989–94. [PubMed: 17618283]
8. 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(8):984–8. [PubMed: 17618284]
9. 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(11):1315–7. [PubMed: 17934461]
10. 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(1):26–8. [PubMed: 18084292]
11. 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;40(5):623–30. [PubMed: 18372905]
12. 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;40(5):631–7. [PubMed: 18372901]

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

13. 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(12):1426–35. [PubMed: 19011631]

14. Amundadottir LT, Sulem P, Gudmundsson J, et al. A common variant associated with prostate cancer in European and African populations. Nat Genet 2006;38(6):652–8. [PubMed: 16682969]

15. Freedman ML, Haiman CA, Patterson N, et al. Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. Proc Natl Acad Sci U S A 2006;103(38):14068–73. [PubMed: 16945910]

16. Haiman C, Patterson N, Freedman ML, et al. Three regions within 8q24 independently modulate risk for prostate cancer. Nature Genet 2007;39:638–44. [PubMed: 17401364]

17. Haiman CA, Le Marchand L, Yamamato J, et al. A common genetic risk factor for colorectal and prostate cancer. Nat Genet 2007;39(8):954–6. [PubMed: 17618282]

18. Yeager M, Xiao N, Hayes RB, et al. Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. Hum Genet 2008;124(2):161–70. [PubMed: 18704501]

19. Poynter JN, Figueiredo JC, Conti DV, et al. Variants on 9p24 and 8q24 are associated with risk of colorectal cancer: results from the Colon Cancer Family Registry. Cancer Res 2007;67(23):11128–32. [PubMed: 18056436]

20. Le Marchand L, Wilkens LR. Design considerations for genomic association studies: importance of gene- environment interactions. Cancer Epidemiol Biomarkers Prev 2008;17(2):263–7. [PubMed: 18268108]

21. Evans DN, Marchini J, Morris AP, Cardon LR. Two-stage two-locus mdels in genome-wide association. PLOS Genet 2006;2:1424–32.

22. Le Marchand L. Combined influence of genetic and dietary factors on colorectal cancer incidence in Japanese Americans. Monograph Natl Cancer Inst 1999;26:101–5.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Table 1

Characteristics of Genome-Wide Association Studies Published as of January 2009 and the Colorectal Cancer Susceptibility Loci Identified

Study Reference	**Genotyping platform (Nb. of SNPs)**	**Sample Size for Stage I**	**Sample Size for Subsequ ent Stages**	**Population**	**SNP ID (minor allele frequency in Europeans)**	**Gene/Region**	**OR per allele**	**p-value over all sample sets**
Zanke Nat Genet 2007 39:989–94. (7)	Illumina and Affymetrix (99,632)	1,257 cases/1,336 Controls[1]	4,024 cases/4,042 controls	First-stage: Canada Other stages: Canada, US, Scotland	rs10505477 (0.50) rs719725	8q24 9p24	1.18 1.14	1.41×10^{-8} 1.32×10^{-5}
Tomlinson Nat Genet 2007 39:984–8. (8)	Illumina (547,647)	930 cases/960 controls[2]	7,334 cases/5,246 controls	First-stage: UK Second stage: UK	rs6983267 (0.49)	8q24	1.21	1.27×10^{-14}
Broderick Nat Genet 2007 39:1315–7. (9)	Affymetrix (550,163)	940 cases/965 controls[2]	7,473 cases/5,984 controls	First-stage: UK Second stage: UK	rs4939827 (0.52)	18q21 *SMAD7*	1.18	1.0×10^{-12}
Jaeger Nat Genet 2008 40:26–8. (10)	Illumina (547,647)	730 cases/960 controls[2]	4,500 cases/3,860 controls	First-stage: UK Second stage: UK	rs4779584 (0.19)	15q13 *CRAC1*	1.26	4.4×10^{-14}
Tomlinson Nat Genet 2008 40:623–30. (11)	Illumina (550,163)	940 cases/965 controls[2]	17,891 cases/17,575 controls	First-stage: UK Second stage: UK, EU	rs10795668 (0.33) rs16892766 (0.07)	10p14 8q23.3 *EIF3H*	0.89 1.25	2.5×10^{-13} 3.3×10^{-18}
Tenesa Nat Genet 2008 40:631–7. (12)	Illumina (541,628)	981 cases/1,002 Controls[3]	16,476 cases/15,351 controls	First-stage: Scotland Second stage and replication: Canada, UK, Israel, Japan, EU	rs4939827 (0.52) rs7014346 (0.37) rs3802842 (0.29)	18q21 *SMAD7* 8q24 11q23	1.20 1.19 1.11	7.8×10^{-28} 8.6×10^{-26} 5.8×10^{-10}
COGENT Nat Genet 2008 40:1426–35 (13)	Multiple (38,710)	6,780 cases, 6,843 controls	13,406 cases, 14,012 controls	Fist-Stage: UK Replication: EU, Canada	rs4444235 (0.46) rs9929218 (0.29) rs10411210 (0.10) rs961253 (0.36)	14q22.2 *BMP4* 16q22.1 *CDH1* 19q13.1 *RHPN2* 20p12.3	1.11 1.20 0.87 1.12	8.1×10^{-10} 1.2×10^{-8} 4.6×10^{-9} 2.0×10^{-10}

[1] population-based controls recruited using random-digit dialing and other population-based assessment lists; matched to cases on age and sex.

[2] controls were spouses or partners of European ancestry (UK resident) unaffected by cancer and without a family history of CRC.

[3] controls were cancer-free and identified from the general population and matched to cases by a