Title

The Hunting of the Snark: Whither Genome-Wide Association Studies for Colorectal Cancer?

Journal

Gastroenterology, 150(7)

ISSN

0016-5085

Authors

Carmona, Luis G Carvajal
Tomlinson, Ian

Publication Date

2016-06-01

DOI

10.1053/j.gastro.2016.04.021

Peer reviewed

The Hunting of the Snark: Whither Genome-Wide Association Studies for Colorectal Cancer?

See "Identification of susceptibility loci and genes for colorectal cancer risk," by Zeng C, Matsuda K, Jia W-H, et al, on page 1633.

In the Lewis Carroll nonsense poem, *The Hunting of the Snark*, a curious assembly of characters with a variety of dubious skills sets forth on a poorly defined quest to find the half-real snark. On the way, the pursuers use a number of approaches, and in turns collaborate, fall out with each other and, in at least one case, go mad. The only seeker who claims to find the snark, disappears.

Genome-wide association studies (GWAS), based on thousands of cases and controls typed at thousands of single nucleotide polymorphisms (SNPs), have identified several variants that associate with gastrointestinal cancer risk. More than 30 colorectal cancer (CRC) predisposition SNPs are known, together with a smaller number of loci for other gastrointestinal cancers.[1–4] It is indisputable, however, that GWAS for other common cancers—notably for breast and prostate—have been much more successful at finding larger numbers of SNPs. It sometimes feels that we will be writing the phrase, "known genetic variants explain only a small fraction of the heritability of gastrointestinal cancers," for many years to come.

The manuscript by Zeng et al[5] in this issue of *Gastroenterology* reports another very useful increase in our CRC genetics knowledge. This study, the largest carried out to date in Asian populations, identified 6 SNPs associated with CRC risk at genome-wide significance ($P < 5 \times 10^{-8}$), including rs4711689 at 6p21, rs2450115, and rs6469656 at 8q23, rs4919687 at 10q24, rs11064437 at 12p13, and rs6061231 at 20q13. The most likely candidate genes affected by the functional variation at each of the 5 sites were respectively reported to be *TFEB* (involved in lysosomal biogenesis), *EIF3H* (initiation of translation), *CYP17A1* (steroid synthesis), *SPS2B2* (proteasome), and *RPS21* (ribosomal biogenesis). Several of these functions seem to be new in terms of CRC pathogenesis. Although most of these SNPs lie in noncoding regions, one of them (rs11064437) has a potential effect on protein sequence, as it falls within the intron 1 splice acceptor of *SPSB2*.

A consideration of 2 of the loci reported by Zeng et al illustrates some of the difficulties in pinning down the functional variation underlying tagSNP signals, especially when comparing ethnic groups.

First, we address the question of how many independent CRC SNPs exist near *EIF3H*? Zeng et al found that 2 SNPs (rs2450115 and rs6469656), near *EIF3H*, mapped to a haplotype block harboring a previously reported CRC SNP in Europeans (rs16892766[6]), which happened to be monomorphic in Asians. Interestingly, rs2450115 and rs6469656, which are in mild linkage disequilibrium (LD) in Asians ($r^2 = 0.20$), remained nominally significant ($P = 9.60 \times 10^{-6}$ for rs2450115 and $P = 8.30 \times 10^{-4}$ for rs6469656) after joint association analysis by Zeng et al. These 2 SNPs were also tested individually in European case-control studies where each was nominally associated with CRC ($P = .0003$ for rs2450115 and $P = .02$ for rs6469656). In Europeans, however, these 2 SNPs have stronger LD ($r^2 = 0.40$) and Zeng et al's joint analysis showed that only rs2450115 remained nominally associated with CRC ($P = .007$), as we ourselves had found in our own previous fine-mapping study.[7] Altogether, these observations are inconclusive, but are consistent with a scenario in which rs2450115 or a strongly correlated SNP is the mostly likely variant driving a single, independent chromosome 8q23 signal.

Second, Zeng et al reported a CRC SNP (rs6061231) mapping to chromosome 20q13, a region containing another SNP (rs4925386) that has previously been associated with CRC in Europeans.[8] These 2 SNPs are in weaker LD in Asians ($r^2 = 0.15$) than in Europeans ($r^2 = 0.44$). After conditional testing in the Asian data sets, only rs6061231 remained significant, naturally leading Zeng et al to suggest that rs6061231 better captured the 20q13 signal. Interestingly, a recent study by Al Tassan,[9] found a third 20q13 CRC variant (rs2427308), which based on 1000 Genomes (1KG) data, is in full LD with rs6061231 in Han Chinese ($r^2 = 1.0$) and in very high LD in East Asians ($r^2 = 0.89$, Figure 1). rs2427308 and rs6061231 also show strong LD in 1000 Genomes Project Europeans ($r^2 = 0.69$, Figure 1). rs6061231 may thus not be an entirely new CRC variant and further studies are needed to assess whether it, rs2427308, and rs4925386 are tagging single or multiple functional 20q13 variants. This work is also important for the detailed functional studies required to determine the identity of the target gene in the region.

Although Zeng et al clearly identified new CRC regions on chromosomes 6p21, 10q24 and 12p13, questions remain about the novelty and number of risk alleles on chromosomes 8q23 and 20q13. Furthermore, although Zeng et al report heterogeneity between Asians and Europeans for 3 SNPs (rs4919687/10q24, rs4711689/6p21, and rs6061231/20q13), it seems unlikely that their preferred explanation, effect allele frequency, is the principal factor causing these differences.

A further inherently troublesome area in GWAS is the identity of the gene(s) which are the targets of the underlying functional variation that influences disease susceptibility. In an attempt to assign genes to SNPs, Zeng et al performed expression quantitative trait locus analysis in anatomically normal colon tissue from 188 Asian patients

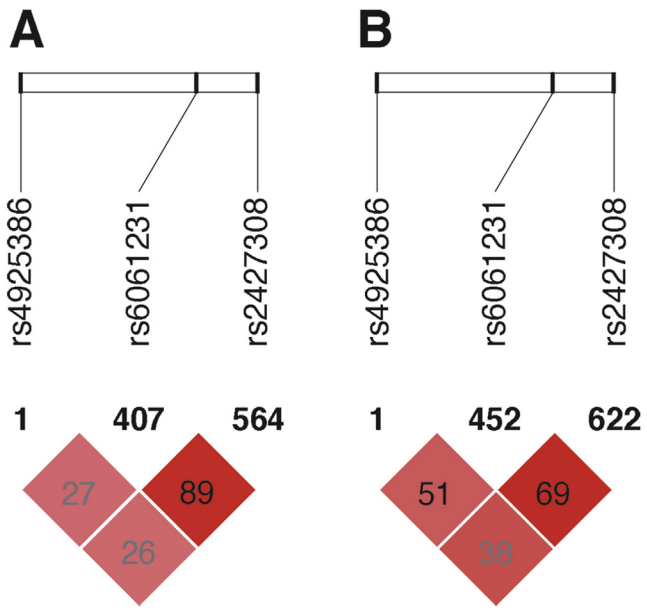

Figure 1. Pairwise linkage disequilibrium (measure as r2) between three colorectal cancer SNPs mapping to chromosome 20q13 in East Asian (A) and (B) European populations from by the 1000 Genomes Project.

with CRC. Of the 5 candidate genes selected because they were nearest to the top SNP in each region, 4—*TFEB, EIF3H, SPSB2* and *RPS21*—were expressed in the colon, and their transcript levels were associated with the nearby CRC SNP genotypes. It is not clear why additional genes in each region were not assessed by expression quantitative trait locus analysis. It would have been interesting, for example, to know whether Zeng et al saw any association of rs6061231 with expression not only of *RPS21*, but also of *LAMA5*, the previously suggested target gene in the region.[8]

Another important issues raised by the Zeng et al study is how to perform cross-ethnicity comparisons. Although cross-ethnic comparisons have been proposed as being very helpful for studies such as fine mapping GWAS signals and examining gene–environment interactions, they have only occasionally been useful for this purpose in practice, owing to the existence of intrinsic problems and uncertainties. For this reason, Zeng et al do not report meta-analyses of Asian and European data sets in their study: although such analyses may increase study power, they risk false positives. Data can be difficult to interpret, given the potential population differences in tagSNP allele frequencies, LD patterns, and effect sizes of the generally unknown functional SNP(s). More generally, GWAS have shown that the effects of functional variation are usually shared across ethnic groups (ie, there is little good evidence of strong genotype–ethnicity interactions). Declaring a SNP specific to Asians or Europeans and vice versa is problematic, because even tens of thousands of cases and controls do not have sufficient power to show that SNPs associated with odds ratios of <1.10 are unambiguously associated (or not associated) with cancer risk. Furthermore, the effect sizes reported when SNPs are discovered are intrinsically likely to overstate the true effects, thus exaggerating differences when those SNPs are tested in different ethnic groups. Zeng et al report Asian SNPs that achieve nominal associations at $P < .05$ in Europeans. This is entirely reasonable, but further cross-ethnic investigations are needed for all SNPs, including those that have been proposed to be associated with CRC in only 1 of the 2 ethnic groups.

Although Zeng et al do not calculate the contribution of the new SNPs to CRC genetics, the increase in the explained heritability is likely to be of the order of, at most, a couple of percent. This is still a nontrivial difference in terms of the explained heritability of CRC. However, it is important to reflect on the fact that breast and prostate cancer SNPs can probably account for a significantly larger fraction of the heritability of those diseases. Can the CRC studies do better?

There are several potential explanations for the "failure" to identify more CRC SNPs. First, CRC has more than its share of high-penetrance genes, 13 at the time of writing (*APC, MUTYH, NTHL1, MSH2, MLH1, MSH6, PMS2, POLE, POLD1, STK11, BMPR1A, SMAD4*, and *GREM1*), with others probably remaining to be found. Breast cancer has 2 such genes (*BRCA1* and *BRCA2*), plus a handful of moderate penetrance genes, and prostate cancer has none. The reasons for this are unclear. Second, CRC has worse survival than either breast or prostate cancer, making patient recruitment more difficult. Based on data from the US Surveillance, Epidemiology and End Results Program (available: http://seer.cancer.gov/statfacts/), the 5-year survival of individuals diagnosed with distant (stages III-IV) breast and prostate tumors (26% and 28%, respectively) doubles that of those diagnosed with distant CRC (13%). Distant tumors, however, represent 20% of all diagnosed CRCs, whereas they represent only 4% and 6% of all prostate and breast tumors, respectively. Third, much of the success of breast and prostate cancer GWAS results from the existence at an early stage of single large consortia, BCAC[10] and PRACTICAL,[11] respectively, that dominated the field. A single, global CRC GWAS consortium has yet to emerge. Fourth, some of the success of BCAC and PRACTICAL is derived from the ability to analyze subgroups of patients, such as those with triple negative or androgen receptor-negative disease respectively. In CRC, the natural equivalent subgroup analysis, of cancer with microsatellite instability, is probably underpowered at the present time, given the fewer CRC patients in the GWAS sample sets and the failure to test for microsatellite instability in most cohorts until recently.

Although we have not found the elusive CRC snark yet, we can anticipate a rush of CRC SNPs when the GAME-ON consortium, an NCI-funded initiative that aims to identify of new CRC loci through meta analyses and extension of GWAS to diverse populations, reports in the near future (available: http://epi.grants.cancer.gov/gameon/index.html). It is hoped that this will enhance global interest in CRC genetics, and in the genetics of other digestive system cancers. Furthermore, an important missing set of studies is GWAS for CRC in African or admixed populations. For some SNPs, factors such as effect allele frequencies, may empower these studies to detect new risk SNPs, which can then be

analyzed in Asian and European populations. Many esophageal and gastrointestinal cancers are amenable to prevention using proven methods that have few complications and side effects. In addition, premalignant lesions can be often removed entirely during screening. It is somewhat ironic that these cancers are lagging behind others when it comes to identifying people in the general population who are at greatest genetic risk. To finish as we started, with a literary allusion, this time from Garcia Marquez's *No One Writes to the Colonel*, we hope that the infinite wait of the colonel for his pension does not presage a similar wait for a full genetic dissection of CRC and other gastrointestinal cancers.

LUIS G. CARVAJAL CARMONA
Genome Center and Department of Biochemistry
and Molecular Medicine
School of Medicine
University of California
Davis, California

IAN TOMLINSON
Oxford Centre for Cancer Gene Research
Wellcome Trust Centre for Human Genetics
University of Oxford
Oxford, United Kingdom

References

1. Carethers JM, Jung BH. Genetics and genetic biomarkers in sporadic colorectal cancer. Gastroenterology 2015;149:1177–1190 e3.
2. Reid BJ, Paulson TG, Li X. Genetic insights in Barrett's esophagus and esophageal adenocarcinoma. Gastroenterology 2015;149:1142–1152 e3.
3. Tan P, Yeoh KG. Genetics and molecular pathogenesis of gastric adenocarcinoma. Gastroenterology 2015;149:1153–1162 e3.
4. Whitcomb DC, Shelton CA, Brand RE. Genetics and genetic testing in pancreatic cancer. Gastroenterology 2015;149:1252–1264 e4.
5. Zeng C, Matsuda K, Jia W-H, et al. Identification of susceptibility loci and genes for colorectal cancer risk. Gastroenterology 2016;150:1633–1645.
6. 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;40:623–630.
7. Carvajal-Carmona LG, Cazier J-B, Jones AM, et al. Fine-mapping of colorectal cancer susceptibility loci at 8q23.3, 16q22.1 and 19q13.11: refinement of association signals and use of in silico analysis to suggest functional variation and unexpected candidate target genes. Hum Mol Genet 2011;20:2879–2888.
8. Houlston RS, Cheadle J, Dobbins SE, et al. Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat Genet 2010;42:973–977.
9. Al-Tassan NA, Whiffin N, Hosking FJ, et al. A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer. Sci Rep 2015;5:10442.
10. Michailidou K, Beesley J, Lindstrom S, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet 2015;47:373–380.
11. Al Olama AA, Kote-Jarai Z, Berndt SI, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet 2014;46:1103–1109.

Reprint requests
Address requests for reprints to: Luis G. Carvajal Carmona, Genome Center and Department of Biochemistry and Molecular Medicine, School of Medicine, University of California, 451 Health Sciences Drive, Davis, California 95616. e-mail: lgcarvajal@ucdavis.edu.

Acknowledgments
We thank Paul Lott for help with Figure 1.

Conflicts of interest
The authors disclose the following: LGCC acknowledges support from the University of California, Davis, the V Foundation Cancer Research, and the National Cancer Institute (Paul Calabresi Career Development Award for Clinical Oncology K12 at UC Davis, award number K12CA138464) of the National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. IT acknowledges support from the Oxford NIHR Comprehensive Biomedical Research Centre and core funding to the Wellcome Trust Centre for Human Genetics from the Wellcome Trust (090532/Z/09/Z).

Most current article

0016-5085/$36.00
http://dx.doi.org/10.1053/j.gastro.2016.04.021