Cancer Genetics 208 (2015) 35–40

Systematic search for rare variants in Finnish early-onset colorectal cancer patients

Tomas Tanskanen [a], Alexandra E. Gylfe [a], Riku Katainen [a], Minna Taipale [b,e], Laura Renkonen-Sinisalo [a,c], Heikki Järvinen [c], Jukka-Pekka Mecklin [d], Jan Böhm [d], Outi Kilpivaara [a], Esa Pitkänen [a], Kimmo Palin [a], Pia Vahteristo [a], Sari Tuupanen [a], Lauri A. Aaltonen [a,*]

[a] *Department of Medical Genetics, Genome-Scale Biology Program, University of Helsinki, Biomedicum, Helsinki, Finland;* [b] *Genome-Scale Biology Program, Institute of Biomedicine, University of Helsinki, Biomedicum, Helsinki, Finland;* [c] *Department of Surgery, Helsinki University Central Hospital, Helsinki, Finland;* [d] *Department of Surgery, Jyväskylä Central Hospital, University of Eastern Finland, Jyväskylä, Finland;* [e] *Science for Life Center, Department of Biosciences and Nutrition, Karolinska Institute, Huddinge, Sweden*

The heritability of colorectal cancer (CRC) is incompletely understood, and the contribution of undiscovered rare variants may be important. In search of rare disease-causing variants, we exome sequenced 22 CRC patients who were diagnosed before the age of 40 years. Exome sequencing data from 95 familial CRC patients were available as a validation set. Cases with known CRC syndromes were excluded. All patients were from Finland, a country known for its genetically homogenous population. We searched for rare nonsynonymous variants with allele frequencies below 0.1% in 3,374 Finnish and 58,112 non-Finnish controls. In addition, homozygous and compound heterozygous variants were studied. No genes with rare loss-of-function variants were present in more than one early-onset CRC patient. Three genes (*ADAMTS4*, *CYTL1*, and *SYNE1*) harbored rare loss-of-function variants in both early-onset and familial CRC cases. Five genes with homozygous variants in early-onset CRC cases were found (*MCTP2*, *ARHGAP12*, *ATM*, *DONSON*, and *ROS1*), including one gene (*MCTP2*) with a homozygous splice site variant. All discovered homozygous variants were exclusive to one early-onset CRC case. Independent replication is required to associate the discovered variants with CRC. These findings, together with a lack of family history in 19 of 22 (86%) early-onset patients, suggest genetic heterogeneity in unexplained early-onset CRC patients, thus emphasizing the requirement for large sample sizes and careful study designs to elucidate the role of rare variants in CRC susceptibility.

Keywords Genetic predisposition to disease, colorectal neoplasms, age of onset, exome sequencing

Colorectal cancer (CRC) accounts for 10% of new cancers worldwide (GLOBOCAN Project, http://globocan.iarc.fr), and its incidence rises rapidly after 45 years of age. The lifetime risk of CRC is approximately 5%, whereas the risk of developing CRC before the age of 40 years is only 0.08% (SEER database, http://seer.cancer.gov). The etiology of most CRCs is complex and multifactorial, involving interplay between multiple genetic and environmental factors. Inherited factors contribute to CRC risk considerably, but a significant fraction of heritability in CRC patients remains incompletely understood (1).

An estimated 5% of CRC patients, including many of those with early-onset disease or multiple affected family members, are highly predisposed to CRC because of rare single-gene defects in *MLH1*, *MSH2*, *MSH6*, *PMS2*, *APC*, *MUTYH*, *SMAD4*, *BMPR1A*, *STK11*/*LKB1*, or *POLE* (2). Because of large effect sizes, family-based linkage analysis was instrumental in mapping these genes. On the other hand, several low-penetrance CRC susceptibility loci have been discovered through genome-wide association studies

Received September 19, 2014; received in revised form December 12, 2014; accepted December 17, 2014.

* Corresponding author.
E-mail address: lauri.aaltonen@helsinki.fi

2210-7762/$ - see front matter © 2015 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cancergen.2014.12.004

(3), but the allelic architectures and causative variants underlying these associations are mostly undefined.

Next generation sequencing (NGS) has uncovered patterns of human genetic variation in unprecedented detail. Because exome sequencing captures a substantial part of functional and disease-causing genomic variation, it is a promising approach to deciphering the role of rare variants in complex disease predisposition (4). Interest in the pathogenic potential of rare variants has emerged from evolutionary theory (5), as well as the fact that much of the heritability of complex diseases remains unexplained (6).

Despite the potential of NGS in the identification of complex trait genes, success has been limited. Part of the reason is genetic heterogeneity, which increases the sample size required for the identification of culprit genes. In this regard, population isolates could offer unique advantages. The population history of Finland has been characterized in detail. Population bottlenecks and genetic founder effects, as well as geographic isolation, have shaped the genetic structure of the population, leading to reduced genetic heterogeneity (7).

Early age of onset is a key feature of hereditary susceptibility to CRC and other common cancers (8), and early-onset CRC patients might be enriched for undiscovered susceptibility variants. In this study, we exome sequenced a discovery set of 22 unselected Finnish CRC patients who were diagnosed before the age of 40 years, and we used exome sequencing data from 95 Finnish familial CRC patients as a validation set. None of the patients displayed known predisposition syndromes that could account for early age of onset or a positive family history. To identify new potential CRC susceptibility genes, we analyzed rare nonsynonymous variants, and considered both dominant and recessive modes of inheritance.

Materials and methods

Samples

We studied a discovery set of 22 nonsyndromic early-onset CRC cases diagnosed before the age of 40 years, and we used 95 familial CRC cases (with at least one affected first-degree relative) as a validation set. Both sample sets were published previously (9,10) and were derived from a series of 1,514 unselected CRC patients, which was collected in nine central hospitals in southern and eastern Finland between May 1994 and June 1998 and in two of these hospitals from year 1998 to present (9,10). The population-based phase of sample collection in nine hospitals contributed 1,042 CRC patients, and 472 additional CRC patient samples were collected in two hospitals after June 1998. All patients gave informed consent to genetic studies on tumor susceptibility, and the study was approved by the appropriate ethics review board.

Both normal and tumor DNA samples were available from each patient. All tumors had been tested for microsatellite instability (MSI), and known CRC susceptibility syndromes had been diagnosed clinically or molecularly. Data on first-degree relatives and their cancer diagnoses had been acquired from official population registries and the Finnish Cancer Registry (9,10). Of 1,514 CRC patients, 38 (2.5%) had been diagnosed before the age of 40 years. Of 38 early-onset CRC patients, 16 (42%) had known genetic CRC susceptibility syndromes, including hereditary non-polyposis colon cancer (12 of 38, 32%), familial adenomatous polyposis (3 of 38, 7.9%), and juvenile polyposis (1 of 38, 2.6%). Based on pathology reports, there was no evidence of inflammatory bowel disease in any of the 38 early-onset patients. Of the 22 early-onset CRC patients with unknown etiology (Table 1), 10 were female (45%), 12 were male (55%), 2 displayed MSI (9.1%), 18 had cancers of the distal colon or rectum (82%), 13 presented with advanced-stage cancer (Dukes stage C or D, 59%), and 3 had a family history of CRC (14%). Median and mean ages of onset were 35.5 and 33.9 years, respectively, ranging from 21−39 years. Germline DNA samples of these 22 nonsyndromic CRC patients were exome sequenced.

Exome sequencing

Exome sequences were captured with the SureSelect Human All Exon Kit v.1 (Agilent Technologies, Santa Clara, CA). Paired-end 75 base pair reads were obtained with an Illumina HiSeq 2000 (Illumina, San Diego, CA). Exome sequencing data quality was confirmed with FastQC (http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc). Reads were mapped to the human reference genome GRCh37 with the Burrows-Wheeler Aligner, v.0.5.9-r16. The Picard MarkDuplicates tool (http://broadinstitute.github.io/picard/) was used to remove duplicate reads. Reads were realigned locally with the Genome Analysis Toolkit IndelRealigner, and single-nucleotide variants and indels were called with the Genome Analysis Toolkit UnifiedGenotyper, v.2.2-16-g9f648cb (https://www.broadinstitute.org/gatk/). Average coverage was 54, and 87% of the targeted regions were covered by more than 10 reads. An in-house developed comparative analysis tool (RikuRator, unpublished) was used to determine allele frequencies in control exomes and to compare variant calls between CRC cases. When relevant, sequencing reads were manually inspected to exclude false-positive variant calls. Loss-of-function (LoF) variants were annotated with the LOFTEE (Loss-Of-Function Transcript Effect Estimator, https://github.com/konradjk/loftee). Accordingly, we excluded ancestral LoF alleles, LoF variants located in the last 5% of the coding region, splice site variants in small introns (<15 base pairs), and LoF variants surrounded by non-canonical splice sites. Functional effects of missense variants were predicted with PolyPhen-2 and SIFT, using the Ensembl Variant Effect Predictor (http://www.ensembl.org).

Exome sequencing controls

Allele frequencies of all variants were determined in 3,374 Finnish and 58,112 non-Finnish control exomes that were publicly available in the Exome Aggregation Consortium (ExAC) database (http://exac.broadinstitute.org). These individuals had been sequenced in various medical and population genetic studies.