www.impactjournals.com/oncotarget/ Oncotarget, 2018, Vol. 9, (No. 13), pp: 11170-11179

Research Paper

# Two novel colorectal cancer risk loci in the region on chromosome 9q22.32

**Jessada Thutkawkorapin[1], Hovsep Mahdessian[1], Tom Barber[2], Simone Picelli[1], Susanna von Holst[1], Johanna Lundin[1], Laura Valle[3], Vinaykumar Kontham[1], Tao Liu[1], Daniel Nilsson[1], Xiang Jiao[1] and Annika Lindblom[1]**

[1]Department of Molecular Medicine and Surgery, Karolinska Institute, Stockholm SE-17176, Sweden

[2]The Ludwig Center and Howard Hughes Medical Institute at the Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins, Baltimore, MD 21231, USA

[3]Hereditary Cancer Program, Catalan Institute of Oncology, IDIBELL and CIBERONC, Barcelona 08908, Spain

***Correspondence to:*** *Annika Lindblom,* ***email:*** *annika.lindblom@ki.se*

***Keywords:*** *familial colorectal cancer; risk haplotype; cancer predisposition; association study; next generation sequencing*

***Received:*** *November 26, 2017* ***Accepted:*** *January 23, 2018* ***Published:*** *January 29, 2018*



## ABSTRACT

**Highly penetrant cancer syndromes account for less than 5% of all cases with familial colorectal cancer (CRC), and other genetic contribution explains the majority of the genetic contribution to CRC. A CRC susceptibility locus on chromosome 9q has been suggested. In this study, families where risk of CRC was linked to the region, were used to search for predisposing mutations in all genes in the region. No disease-causing mutation was found. Next, haplotype association studies were performed in the region, comparing Swedish CRC cases (2664) and controls (4782). Two overlapping haplotypes were suggested. One 10-SNP haplotype was indicated in familial CRC (OR 1.4, *p* = 0.00005) and one 25-SNP haplotype was indicated in sporadic CRC (OR 2.2, *p* = 0.0000012). The allele frequencies of the 10-SNP and the 25-SNP haplotypes were 13.7% and 2.5% respectively and both included one RNA, *RP11-332M4.1* and *RP11-l80l4.2*, in the non-overlapping regions. The sporadic 25-SNP haplotype could not be studied further, but the familial 10-SNP haplotype was analyzed in 61 additional CRC families, and 6 of them were informative for all markers and had the risk haplotype. Targeted sequencing of the 10-SNP region in the linked families identified one variant in *RP11-332M4.1*, suggestive to confer the increased CRC risk on this haplotype. Our results support the presence of two loci at 9q22.32, each with one RNA as the putative cause of increased CRC risk. These RNAs could exert their effect through the same, or different, genes/pathways, possibly through the regulation of neighboring genes, such as *PTCH1, FANCC, DKFZP434H0512, ERCC6L2* or the processed transcript *LINC00046*.**

## INTRODUCTION

It has been estimated that approximately 13% of all colorectal cancers (CRC) may be due to genetic factors [1]. However, the known predisposing inherited polyposis- and non-polyposis syndromes with highly penetrant mutations in *APC*, *MUTYH*, the DNA mismatch repair (MMR) genes and other even more rare genes, account for less than 5% of all cases [2]. Hypothetically, other high-risk and low-risk genes would explain the majority of the genetic contribution to CRC. Genome-wide linkage analysis (GWL) in CRC families has traditionally been used to identify high-penetrant genes, while genome-wide association studies (GWAS) in CRC patients and controls have been used to find alleles associated with a low/modest risk. Several candidate regions linked to CRC predisposition have been identified by GWL, however, yet no new disease gene or syndrome has been identified

in those genomic regions [3, 4]. In the last years, GWAS have identified alleles associated with a small increased CRC risk, altogether considered to contribute to a minor part of the missing heritability in CRC [5]. At the same time, whole-exome sequencing studies in CRC families have revealed novel rare candidate genes for hereditary CRC, with more or less evidence of causality [6–11].

A sib pair study identified a CRC-associated region, 9q22.2-31.2, subsequently confirmed by analyses in CRC families [12–14]. Genes located within the region, such as *GALNT12*, *AXIN2* or *TGFBR1*, have been suggested as potential causal candidate genes [15–19]. Interestingly, our group could find support for the same locus in a linkage study in a large Swedish family (No. 24) with one individual affected with early onset rectal cancer and several relatives with adenomas (LOD = 2.4) [20]. A subsequent linkage study carried out by our group in 600 individuals from 121 non-FAP/non-LS families identified the exact same locus, as the second-best hit, although still not statistically significant (HLOD = 2.2) [4]. These results prompted further studies in an attempt to define the disease-causing mechanism within this locus in family No. 24 and other families that showed linkage to the same 9q region.

## RESULTS

After the first [20] and before the second [4] linkage study, all exons and exon-intron boundaries of all coding genes within the linked region were (Sanger) sequenced in two cancer/adenoma-affected members (Co-648 and Co-166) of Family No. 24 and no clear deleterious mutation was found (data now shown). All missense variants identified (Supplementary Table 1) were assessed by using association studies that included up to 400 CRC cases and controls, finding no clear association with the disease.

Allele-specific expression (ASE) of TGFBR1, located within the region of interest, was reported to be associated with an increased risk of CRC [19]. However, no TGFBR1 allelic expression imbalance was identified in RNA extracted from peripheral blood lymphocytes of the affected members of Family No. 24 (data not shown). Exon-targeted deletion/duplication analysis using a custom array-CGH design showed no pathogenic or likely pathogenic deletions or duplications in the 9q region in the two affected relatives of Family No. 24 (data not shown). Since the distance between the probes in the array was approximately 2.5Kb, the method was sensitive and accurate enough to capture any deletions or duplications larger than approximately 2.5Kb in the region.

After the second linkage study, whole-exome sequencing was carried out in two affected members of Family No. 24 (Co-166 and Co-213) and in 16 affected members from eight families (No. 8, 13, 275, 296, 350, 478, 740 and 918), which had contributed the most to the HLOD score >2 in the subset of 27 high-risk families studied [4].

Data from whole-exome sequencing for the region of interest were merged to one data set for analysis. After filtering, all exonic non-synonymous variants with a population MAF<20% (source: ExAC; http://exac.broadinstitute.org/) that were shared by the two affected relatives from Family No. 24 were selected. We next searched the other families for mutations in the selected genes. The only gene, which involved Family No. 24 and at least one other family, was *GRIN3A,* where 3 variants were identified: rs62000403, rs3739722 and rs10989563 (Supplementary Table 2). Genotyping of the three *GRIN3A* variants was carried out in 768 familial CRC cases and 768 controls. None of the three variants was associated with the disease (*p*-values: 0.3933, 0.1926 and 0.1840, respectively).

When looking for mutations in the other families only, one gene, *NUTM2G*, displayed variants in the 9q22 linked families. Three heterozygous *NUTM2G* missense variants were identified. Two families (275 and 296), carried rs201544487, one family (296) rs2296815 and a third family (918) rs7866127. All three variants were common in the European population (MAF range: 4.8–12.5%; source ExAC) and were predicted to be neutral by at least out of the *in silico* predictors used, suggesting a non-pathogenic nature (Supplementary Table 2).

The absence of suggestive deleterious mutations within the coding region included in the region of interest, led us to hypothesize that the region might hold a genetic risk factor within the non-exonic regions. Moreover, the presence of the 9q22 linkage in a total of Swedish families prompted us to test the hypothesis of a Swedish founder haplotype. Next, we performed a haplotype association study using 2664 consecutive CRC cases and 4782 controls from an ongoing GWAS (CORECT).

Genotypes for 500 markers in the region (rs16909975–rs12237372) were accessed and two windows (10 and 25) were studied, thus requiring a *p*-value lower than 0.00005 for statistical significance. One suggested risk haplotype was a 25-SNP haplotype with an OR of 1.8 and a *p*-value 0.000058 (borderline statistically significant) and a haplotype frequency of 2.7% in the normal population. To find out if known CRC families carried this haplotype, all 25 markers were genotyped in a separate set of 61 familial CRC cases and their relatives, to find out if any of those families could have the suggested haplotype. None of the 61 familial CRC haplotypes matched the 25 markers on the haplotype even when considering those not fully informative for all 25 SNPs. The cases in the association studies were consecutive cases, and 82% were sporadic. We hypothesized that perhaps the haplotype would be less prevalent among the familial cases to explain why we could not see this risk haplotype among our 61 familial cases. The results from single SNP analysis supported this hypothesis, since the SNP with the best *p*-value, rs6477733 ($p = 0.00019$, Supplementary Figure 1), 1Mb from the haplotype, was more prevalent in sporadic cases

(3%) compared to familial (1%), suggesting a difference between familial and sporadic cases.

To test this hypothesis, the 2664 CRC cases were split into 481 familial (those with at least one other CRC case in their family) and 2183 sporadic cases. The analysis was repeated, again using two windows (10 and 25), and this time familial cases and sporadic cases were analyzed separately, but using the same controls for both analysis. As a result, from the analysis of the sporadic cases the same 25-SNP haplotype was found with improved statistical significance (OR = 2.2, $P$ = 0.0000012, haplotype frequency in normal controls 2.5%), confirming our hypothesis (Figure 1). The haplotype frequency of this 25-SNP haplotype in the 481 familial cases was estimated to be 1.8%, consistent with lack of this haplotype among the 61 familial cases. In the analysis using only familial cases, a different 10-SNP haplotype was suggested with an OR = 1.4 ($p$ = 0.000093), and a haplotype frequency in normal controls of 13.7% (Figure 1). The frequency of this haplotype among sporadic cases was similar to controls, 14%, while the haplotype frequency in familial was 18%. The two haplotypes were overlapping for 6 markers (rs7854560, rs7860540, rs930280, rs6478302, rs109894996 and rs10989747). Since the 25-SNP haplotype was already tested and did not segregate within the separate 61 familial cases, we now tested the markers for the 10-SNP haplotype in the same 61 families, which included eight of the families from the linkage studies above. The full 10-SNP haplotype was found in one of the linked families, No. 24, and in five other families (254, 325, 340, 415, 485), as well as suggested (although not fully informative for all markers) in two of the other linked families (350, 740) plus 13 additional families (12, 26, 60, 70, 161, 288, 309, 310, 409, 425, 470, 660, 1085), in total 21 families (17%) corresponding well with the estimated haplotype frequency (18%) from the association study (Figure 2). These results confirm a 10-SNP-founder haplotype among Swedish familial cases and suggest a 25-SNP haplotype in Swedish sporadic CRC cases.

The region outlined by these two haplotypes, rs6478058-rs17393861, is in the intergenetic region between *PTCH1* and *LINC00046* (Figure 1). This region harbors two lincRNAs, *RP11-332M4.1* and *RP11-l80l4.2*. *RP11-332M4.1* is located within the non-overlapping part of the 10-SNP haplotype, while *RP11-l80l4.2* is located within the non-overlapping part of the 25-SNP haplotype (Figure 1).

Targeted sequencing of the whole region, suggested by our linkage study, was performed in samples from 46 families, including Family No. 24 and from families (254, 325, 340, 415) with the complete 10-SNP haplotype, and another 9 from 15 families with the suggested haplotype (26, 60, 70, 161, 288, 309, 425, 740, 1085). Since family No. 24 had most support for a genetic predisposition, based on both highest LOD and a fully informative 10-SNP-risk haplotype, two affected cousins (Co-213 and Co-166) from this family were selected for sequence analysis within the 10-SNP-risk haplotype (Supplementary Table 3). Since the association study suggested this haplotype to be present in 14% of the normal population, candidate variants with a population MAF<25% were selected for further analyses (Supplementary Table 3).

Nine variants were considered artifacts related to the difficulties for the annotation program to accurate interpret repeats. Two SNPs, rs3215956 and rs199596284, were ruled out as they showed the same frequency in 96 cases and 96 controls that were Sanger sequenced. Five variants in the two affected cousins in Family No. 24 remained as potential candidates in familial CRC cases. First, rs34556283, within *RP11-332M4.1*, with a population MAF of 17%, was also identified in two families with a complete haplotype (254 and 325), and in seven (26, 60, 161, 288, 425, 740, 1085) of the nine families with a suggested haplotype (Supplementary Table 3). Second, four SNPs, rs34227262, rs13301752, rs7024435 and rs7036222, (population MAF 19%) were located within the risk haplotype, but outside the lincRNA *RP11-332M4.1*. They were present in all four (254, 325, 340, 415) families with the complete haplotype, and in seven (70, 161, 288, 309, 425, 740, 1085) of the nine families suggested to have the haplotype (Supplementary Table 3). The rs34556283 variant was genotyped in 725 consecutive CRC cases and 671 controls, showing a difference between cases and controls with an OR similar to the OR from the haplotype analysis (18.5% in cases and 16.4% in controls; OR = 1.15; $p$ = 0.2 n.s.). Testing one (rs7024435) of the four variants on the same haplotype, did not show any difference between cases and controls when genotyped in 320 cases and 341 controls (20.5% in cases and 20.7% in controls; OR = 0.99; $p$ = 0.29 n.s.).

## DISCUSSION

The CRC candidate region on 9q22 has been suggested by several studies [12–14], although not in any previous CRC GWAS, which is surprising considering the relatively high OR (2.2) in sporadic cases in the present study. We think this might be explained by the fact that we did haplotype analysis rather than single SNP analysis. The best $p$-value in the single-SNP analysis was much less significant (0.00019) compared with our first haplotype analysis (0.000058). The results are consistent with what we found in our previous haplotype analysis [21], (Oncotarget, in press). Besides, this region holds also other known CRC genes, which could have influenced results from single-SNP GWAS [15–19].

The background for this study was the repeated findings suggesting a CRC susceptibility locus on chromosome 9q22. First, it was suggested by a sib-pair study [14], then was confirmed in familial CRC [13] and by us in a follow-up study in family No. 24 [20]. This

family was one of the families contributing to a suggested locus in the 9q region already in an earlier linkage study in Finnish, Danish and Swedish families by Päivi Peltomäki (unpublished data), but when the sib-pair study was published, family No. 24 was extended to include more family members, and the published locus on 9q could be confirmed [20]. Still, no mutation was detected in the family using Sanger sequencing of all genes in the region (Bert Vogelstein, unpublished data).

The exact same locus came up again as a result in our recent linkage study of 126 families [4]. Thus, we decided to continue the search for genes in the region, now including these new families. Whole exome analysis in members of family No. 24 and other linked families did not find any support for a causative gene in the region. Since the suggestion for an increased risk came from both high-risk families [4, 13] and low/moderate risk families [14, 20], we decided to use an approach of haplotype analysis to search for a founder cause.

Data from an ongoing GWAS in CORECT, a consortium for association studies in CRC, was used to study this 9q region. The results suggested two

| | | | Familial cases | Sporadic cases | |
|---|---|---|---|---|---|
| | | | 9.3E-05 | 1.2E-06 | P-value |
| | | | 1.4149 | 2.1793 | OR |
| | | Chrom9 | 0.1843 | 0.05635 | Freq. in cases |
| | SNP | location | 0.1377 | 0.02667 | Freq. in controls |
| | rs6478058 | 98360213 | | | |
| * | rs928618 | 98362492 | A ← rs34556283 | | |
| | rs10120219 | 98364547 | G ← rs34117262, rs13301752, rs7024435 | | |
| | rs4743090 | 98376367 | G ← rs7036222 | | |
| | rs7864457 | 98381360 | G | | |
| | rs7854560 | 98382950 | G | G | |
| | rs7860540 | 98383097 | G | G | |
| | rs930280 | 98391111 | A | A | |
| | rs6478302 | 98392340 | A | A | |
| | rs10989496 | 98397006 | G | G | |
| | rs10989747 | 98402621 | G | G | |
| | rs9695781 | 98407526 | | A | |
| | rs1582073 | 98420881 | | G | |
| | rs354276 | 98439240 | | A | |
| ** | rs915228 | 98444303 | | A | |
| | rs458477 | 98445060 | | A | |
| | rs817184 | 98447172 | | C | |
| | rs460175 | 98478118 | | A | |
| | rs1967908 | 98479529 | | G | |
| | rs354271 | 98483660 | | A | |
| | rs2119 | 98489142 | | A | |
| | rs16910200 | 98501892 | | A | |
| | rs7035549 | 98502921 | | G | |
| | rs700966 | 98507588 | | G | |
| | rs4743271 | 98508628 | | A | |
| | rs4742743 | 98515578 | | A | |
| | rs10981787 | 98516889 | | G | |
| | rs4743306 | 98534372 | | G | |
| | rs7035466 | 98535990 | | A | |
| | rs1836404 | 98550264 | | A | |
| | rs17393861 | 98554957 | | | |

**Figure 1: Haplotypes revealed in association studies.** *RNA RP11-332M4.1; **RNA RP11-180l4.2.

separate risk factors with one haplotype each. When we analyzed the samples using sporadic and familial samples separately, we had support for our hypothesis of two founder effects. In support of the results, to 21 (all relatives were not fully informative for all 10 SNPs) of the CRC families included in the current study had the familial 10-SNP haplotype. None of the tested familial cases had the 25-SNP haplotype, suggested as risk factor in sporadic CRC, which was surprising but consistent with the low frequency among familial cases. The 25-SNP haplotype in the sporadic cases had an OR 2.2, while the 10-SNP haplotype in the familial cohort had an OR of 1.4. The relatively low OR in the familial cohort suggested a modifier role, probably exerting its effect together with other risk factors as expected in complex diseases, rather than as a high-risk gene. It was not possible to study haplotypes for sporadic cases, since no family members were collected in the Swedish Low Risk Study, which recruited consecutive CRC cases. Haplotypes could be studied in families, where both cases and relatives were recruited, when they were undergoing genetic counseling in. Most important, family No. 24, showed the full 10-SNP haplotype. Analysis of sequencing data for the region suggested one SNP, rs34556283, to possibly be the disease-causing variant within the RNA *RP11-332M4.1*.

The results suggested two risk loci, one in familial and one in sporadic CRC. Although it cannot be excluded that they both target the same risk locus, we think this is unlikely, since the support for the 25-SNP haplotype was stronger when the familial samples were removed. It is possible that both these loci, each with its own RNA, hold risk factors with a somewhat different effect on their own, or together with other genetic risk factors, and that the respective RNA is the target for the mutations. *RP11-332M4.1* and *RP11-l80l4.2* are long intergenic non-coding RNAs. They were manually annotated in the VEGA database [22] as part of the ENCODE project [23]. They are still poorly understood. LincRNAs has been suggested to be able to reprogram chromatin state as well as being involved in transcriptional silencing during cancer development [24–27]. The effect of mutations could relate to neighboring genes, such as the *PTCH1 or FANCC* gene or a processed transcript *LINC00046,* a protein coding gene *DKFZP434H0512* or *ERCC6L2*. The *PTCH1* gene is a well-known cancer gene involved in predisposition to basal cell carcinoma and other human tumors, but has also been implicated in CRC [28–30]. The *FANCC* gene is also well known to predispose to cancer and was recently also suggested in CRC [31]. *ERCC6L2* belongs to a family of helicases related to yeast *Snf2*, and mutations have been implicated in DNA repair and mitochondrial function [32]. A previous study also used a haplotype approach, in familial samples and could define two regions, both close but proximal to our region [12].

Even if Sweden today is not a very homogenous population, it was more so when the CRC patients were born, and our study demonstrates how novel risk factors can be found in such a population using haplotype analysis. It also demonstrates how linkage analysis not only can be used to find high-penetrant susceptibility loci, but also low-risk variants involved in complex disease. The difficulties to define a genetic variant outside the exome are obvious. Here, at least one variant was suggested, but it cannot be ruled out that limitations in current status of NGS have hidden other possible variants. Furthermore, it will be challenging to demonstrate the effect of a specific non-exonic variant.

We conclude that this study suggested two different risk alleles within the 9q22 locus. One, involving the RNA *RP11-l80l4.2,* was suggested in sporadic CRC (OR 2.2) and the other involving the RNA *RP11-332M4.1* in familial CRC (OR 1.4) suggesting the latter to act as a modifier or in complex inheritance with other genetic risk factors. Further studies will show how the risk alleles at this risk locus on 9q22 influence the risk of CRC.

| | Chrom 9 | Best 10-SNP | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | Family no: | | | | | | | | | | | | | | | | | | | | |
| SNP | location | Haplotype | 24 | 254 | 325 | 340 | 415 | 485 | 12 | 26 | 60 | 70 | 161 | 288 | 309 | 310 | 350 | 409 | 425 | 470 | 660 | 740 | 1085 |
| rs928618 | 98362492 | A | A | A | A | A | A | A | A | A | A | A | A | A | A | na | A | A | na | na | A | na | na |
| rs10120219 | 98364547 | G | G | G | G | G | G | G | G | G | G | G | G | G | G | na | G | G | G | G | na | G | G |
| rs4743090 | 98376367 | G | G | G | G | G | G | G | G | G | G | G | G | G | na | G | G | G | G | G | G | G | G |
| rs7864457 | 98381360 | G | G | G | G | G | G | G | G | G | G | G | G | G | na | G | G | G | G | G | na | G | G |
| rs7854560 | 98382950 | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | na | na | na | G |
| rs7860540 | 98383097 | G | G | G | G | G | G | G | na | na | na | na | na | na | G | na | na | G | na | G | G | G | na |
| rs930280 | 98391111 | A | A | A | A | A | A | A | na | A | na | na | na | na | A | na | na | A | A | A | na | A | na |
| rs6478302 | 98392340 | A | A | A | A | A | A | A | na | A | na | na | na | na | A | na | na | na | A | A | A | A | na |
| rs10989496 | 98397006 | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | G | na | G | G |
| rs10989747 | 98402621 | G | G | G | G | G | G | G | na | G | na | G | na | na | G | na | na | na | G | na | na | na | na |
| | | | * | * | ** | | | | | * | * | | * | * | | | | | | * | | * | * |
| | | | § | §§ | §§ | § | § | | | | | § | § | § | § | | | | § | | | § | § |
| | | | # | | | | | | | | | | | | | | # | | | | | # | |
| | | | % | % | % | % | % | | | % | % | % | % | % | % | | | | % | | | % | % |

**Figure 2: Families with the full or incomplete 10-SNP haplotype.** §,§§, heterozygous and homozygous for rs34117262, rs13301752, rs7024435, rs7036222; *,**, heterozygous and homozygous for rs34556283; na, not available; #, families linked to the region; %, families with targeted sequencing data.

## MATERIALS AND METHODS

### Swedish study participants

#### Familial cases used for sequencing- and haplotype analysis:

Familial cases were defined as coming from families where at least two first or second-degree relatives were affected with CRC. Family No. 24 was described in [20]. The families from the second linkage study were described in [4]. In total, whole-exome sequencing was performed in 98 familial CRC cases, which included family members from the families linked to the region. All CRC families were recruited through the Department of Clinical Genetics, Karolinska University Hospital Solna (Sweden). All families had undergone a full genetic investigation, and FAP and Lynch syndrome were excluded in all families using current clinical routines [33]. Two family members from each of a total of 61 CRC families were interrogated for the specific haplotypes. One case and one parent or child were genome-wide genotyped in order to analyze the 61 haplotypes to search for any candidate risk haplotype resulting from our studies.

#### CRC patients and controls used for association studies

The genotyping data used for the association haplotype study of the region, was obtained from CRC patients recruited in a nationwide study, the Swedish Low-risk Colorectal Cancer Study. The cases were from a cohort of more than 3300 consecutive CRC patients from 14 hospitals in and around Stockholm and Uppsala between 2004 and 2009, and gave informed consent and blood for genetic studies. All cases were interviewed, by the same person, about their family history of CRC and other malignancies. Cancer in first- and second-degree relatives and cousins was recorded, and pedigrees for the families of the index-person (the patient) were constructed. All diagnoses in family members, which could have been CRC were verified using medical records or death certificates. Other diagnoses were coded as stated by the index case. Cases with no relatives diagnosed with CRC were considered sporadic. Familial CRC was defined as cases with at least one relative with CRC in the family as defined above. All patients where relatives were at increased risk because of the family history were offered genetic counselling. Sex, age and tumor location of the index-patients were recorded based on medical records. Tumors were assigned locations in caecum, ascending colon, hepatic flexure, transverse colon, splenic flexure, descending colon, sigmoid colon or rectum. All tumors underwent evaluation directly after surgery by a local pathologist. The tumors were staged according to both AJCC classification and TNM system. From all patients in that study, detailed pedigrees were obtained to be able to classify each case as familial or, mostly, sporadic. As controls were used 4782 healthy unrelated twins from the Swedish Twin registry [34].

### Array-CGH

A custom designed array-CGH analysis was used for exon targeted detection of deletions and duplications in the 9q region. Agilent Technologies SureDesign was used to design the targeted 4x180K array (Oxford Gene Technologies, Oxfordshire, UK). This design has 8908 probes targeting the 9q region with a median probe spacing of 818 base-pairs giving a resolution using a 3probe cut-off of about 2.5 Kb. Experiments were performed at the Department of Clinical Genetics at Karolinska University Hospital, Stockholm, Sweden according to the manufacturer's protocol. Slides were scanned using the Agilent Microarray Scanner (G2505C, Agilent technologies, USA). Raw data were normalized using Feature Extraction Software (10.7.3.1, Agilent Technologies, USA), and log2 ratios were calculated by dividing the normalized intensity in the sample by the mean intensity across the reference sample. The log2 ratios were plotted and segmented by circular binary segmentation in the CytoSure Interpret software (Oxford Gene Technology, Oxfordshire, UK). Oligonucleotide probe positions were annotated to the human genome assembly hg19 (www.genome.ucsc.edu).

### Sanger sequencing

Sanger sequencing was performed as previously described [35, 36]. Primer pairs were designed to amplify the coding regions of all genes in the 9q region. PCR products were purified using Agencourt AMPure Beads (Beckman Coulter) and sequenced with nested PCR primers. Sanger sequencing data was analyzed as previously [35].

### Exome sequencing of germline DNA from 98 familial CRC cases

DNA was quantified using a Qubit Flurometer (Life Technologies). Sequencing libraries were prepared according to the TruSeq DNA Sample Preparation Kit EUC 15005180 or EUC 15026489 (Illumina). Briefly, 1–1.5 ug of genomic DNA was fragmented using a Covaris (Covaris, Inc.). Thirty-seven of the DNA samples were fragmented according to the Covaris 400 bp protocol and 61 samples were fragmented according to the SureSelect Protocol. After fragmentation, all samples were subjected to end-repair, A-tailing, and adaptor ligation of Illumina Multiplexing PE adaptors. An additional gel-based size selection step was performed for the 37 samples. The adapter-ligated fragments were subsequently enriched by PCR followed by purification using Agencourt AMPure Beads (Beckman

Coulter). Exome capture was performed by pre-pooling equimolar amounts and performing enrichment in 5- or 6-plex reactions according to the TruSeq Exome Enrichment Kit Protocol (EUC 15013230). Library size was checked on a Bioanalyzer High Sensitivity DNA chip (Agilent Technologies) while concentration was calculated by quantitative PCR. The pooled DNA libraries were clustered on a cBot instrument (Illumina) using the TruSeq PE Cluster Kit v3. Paired-end sequencing was performed for 100 cycles using a HiSeq 2000 instrument (Illumina) with TruSeq SBS Chemistry v3, according to the manufacturer's protocol. Base calling was performed with RTA (1.12.4.2 or 1.13.48) and the resulting BCL files were filtered, de-multiplexed, and converted to FASTQ format using CASAVA 1.7 or 1.8 (Illumina). Data have been analyzed using the bcbb package (https://github.com/bbcb). After sequencing, the samples have been aligned to the reference genome hg19GRCh37 using BWA [37], sorted and PCR duplicates were removed with Picard (http://broadinstitute.github.io/picard/). The calculation of mapping and enrichment statistics were done with Picard and GATK. Variants were called using GATK and followed a best practice procedure implemented at the Broad Institute [38].

### Mutation annotation

The output mutations in variant call format (VCF) were annotated using ANNOVAR [39], which generated an excel-compatible file with gene annotation, amino acid change annotation, dbSNP identifiers [40], and 1000 Genomes Project allele frequencies [41].

### Genotyping and quality control of the association study

DNA was extracted from peripheral blood samples for both the cases and the controls. The 2690 cases were genotyped at the Center for Inherited Disease Research at Johns Hopkins University, US, using the Illumina Infinium® OncoArray-500K BeadChips. The 4782 controls from the Swedish TwinGene registry were genotyped in Uppsala, Sweden using the Illumina OmniExpress BeadChips. The twin cohort and the Colorectal Cancer Transdisciplinary Study (CORECT) cohort went through quality control (QC) at their corresponding genotyping centers. In total 240370 SNPs were shared between the two platforms on which the data was merged and TOP strand format was accounted for. 9117 (2690 cases and 6427 controls) individuals were proceeded for QC analysis. In the first QC round (QC1), heterozygous haploid genotypes were excluded as well as samples with gender inconsistency and same position variants. The 239113 SNPs and 9114 individuals (2688 cases and 6426 controls) passed QC1. A second QC stage (QC 2) was performed on the merged data, where SNPs with <98% call rate, <1% minor allele frequency (MAF) and those inconsistent with Hardy–Weinberg (hwe 0.0001) equilibrium in controls were removed. 223065 SNPs remained after QC 2. In the third and final QC (QC 3) a multidimensional scaling (MDS) analysis was conducted on all the remaining markers for the purpose of population stratification and to identify ethnic outliers. The outliers were excluded from the dataset while the rest were plotted in an MDS plot (Supplementary Figure 1). After QC 3, 223065 SNPs and 9068 individuals (2664 cases, 6408 controls) remained to perform further downstream analyses.

### Genome-wide association study

Haplotype association studies were performed using PLINK V1.07 [42] on three sub-groups of CORECT genotyping data, familial ($n$ = 481), sporadic ($n$ = 2183), and familial + sporadic ($n$ = 2664) as cases, and Swedish Twin Registry [34] as controls.

### Genotyping of familial samples for testing of haplotypes

Genomic DNA was extracted from peripheral blood using standard procedures. Genotyping of in total 587 individuals, familial CRC cases and their relatives, was performed using the Illumina HumanOmniExpress-12v1_H BeadChip. The results, 730,525 SNPs, were analyzed using the software GenomeStudio 2011.1 from Illumina Inc. Average sample call rate per SNP with sample call rate >0 was >99% and the overall reproducibility >99.99%. Arrays were processed according to manufactures protocol at the SNP&SEQ Technology Platform at Uppsala University and available on request (www.genotyping.se).

### Targeted sequencing of the 9q region

Capture sequencing of 46 familial CRC patients was performed by Axeq Technologies, US, using a SureSelect target enrichment system process followed by 100 bp paired-end sequencing on an Illumina HiSeq2000 sequencer. After sequencing, bioinformatics analysis of the FASTQ files included alignment of sequence reads to the reference human genome (GRCh37/hg19) using BWA and SAMTools, applying GATK [38, 43, 44] base quality score recalibration, indel realignment, duplicate removal, variant calling and annotation (dbSNP and 1000 Genome Project).

### Association studies of missense mutations

Association studies were performed using Taqman SNP Genotyping Assay (Thermo Fisher Scientific).

## Ethics

All patients gave written informed consents in accordance with Swedish legislation (2003:460) and

the study was approved by the Regional Research Ethics Committee, Dnr: 2002-20489, 2008/125-2031/2, 2014/1324-31 and 2016/24-31/1.

## Abbreviations

SNP: single nucleotide polymorphism; LOD: logarithm of odds; HLOD: heterogeneity LOD; OR: odds ratio; MAF: minor allele frequency; Mb: megabase.

## Author contributions

JT analyzed targeted and exome sequencing results, performed haplotype association studies and participated in writing the draft. HM participated in quality control of genotyping data. TB performed Sanger sequencing. SP and SvH performed linkage analysis. JL performed array-CGH test. LV performed allele specific expression analysis. VK, TL and XJ performed fine-mapping analysis. DN provided bioinformatics support. AL conceptualized and designed the study, was in charge of acquisition of data, discussion of the results, writing and revising the manuscript and took overall coordination and responsibility of the study.

## ACKNOWLEDGMENTS

We thank all patients for their contribution. We acknowledge Albert de la Chapelle and Bert Vogelstein for their valuable contribution. We acknowledge The Swedish Twin Registry for access to data. Genotyping was performed by the SNP&SEQ Technology Platform in Uppsala, which is supported by Uppsala University, Uppsala University Hospital, Science for Life Laboratory – Uppsala and the Swedish Research Council (Contracts 80576801 and 70374401).

Colorectal Transdisciplinary Study (CORECT): The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the CORECT Consortium, nor does mention of trade names, commercial products or organizations imply endorsement by the US Government or the CORECT Consortium.

## CONFLICTS OF INTEREST

The authors declare that they have no competing interests.

## FUNDING

The Swedish Study was supported by grants from the Swedish research council; K2015-55X-22674-01-4, K2008-55X-20157-03-3, K2006-72X-20157-01- 2, The Swedish Cancer Society 160458, The Stockholm Cancer Society 161183, and the Stockholm County Council (ALF projects).

CORECT: The CORECT Study was supported by the National Cancer Institute, National Institutes of Health (NCI/NIH), U.S. Department of Health and Human Services (grant numbers U19 CA148107, R01 CA81488, P30 CA014089, R01 CA197350; P01 CA196569; R01 CA201407).

The Swedish Twin Registry is managed by Karolinska Institutet and receives funding through the Swedish Research Council under the grant no 2017-00641.

## Editorial note

This paper has been accepted based in part on peer-review conducted by another journal and the authors' response and revisions as well as expedited peer-review in Oncotarget.

## REFERENCES

1. Frank C, Sundquist J, Yu H, Hemminki A, Hemminki K. Concordant and discordant familial cancer: Familial risks, proportions and population impact. Int J Cancer. 2017; 140:1510–6. https://doi.org/10.1002/ijc.30583.

2. Aaltonen L, Johns L, Jarvinen H, Mecklin JP, Houlston R. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. Clin Cancer Res. 2007; 13:356–61. https://doi.org/10.1158/1078-0432.CCR-06-1256.

3. Cicek MS, Cunningham JM, Fridley BL, Serie DJ, Bamlet WR, Diergaarde B, Haile RW, Le Marchand L, Krontiris TG, Younghusband HB, Gallinger S, Newcomb PA, Hopper JL, et al. Colorectal cancer linkage on chromosomes 4q21, 8q13, 12q24, and 15q22. PLoS One. 2012; 7:e38175. https://doi.org/10.1371/journal.pone.0038175.

4. Kontham V, von Holst S, Lindblom A. Linkage Analysis in Familial Non-Lynch Syndrome Colorectal Cancer Families from Sweden. Plos One. 2013; 8:7. https://doi.org/10.1371/journal.pone.0083936.

5. Whiffin N, Hosking FJ, Farrington SM, Palles C, Dobbins SE, Zgaga L, Lloyd A, Kinnersley B, Gorman M, Tenesa A, Broderick P, Wang Y, Barclay E, et al. Identification of susceptibility loci for colorectal cancer in a genome-wide meta-analysis. Hum Mol Genet. 2014; 23:4729–37. https://doi.org/10.1093/hmg/ddu177.

6. Nieminen TT, Abdel-Rahman WM, Ristimaki A, Lappalainen M, Lahermo P, Mecklin JP, Jarvinen HJ, Peltomaki P. BMPR1A mutations in hereditary nonpolyposis colorectal cancer without mismatch repair deficiency. Gastroenterology. 2011; 141:e23-6. https://doi.org/10.1053/j.gastro.2011.03.063.

7. Nieminen TT, O'Donohue MF, Wu YP, Lohi H, Scherer SW, Paterson AD, Ellonen P, Abdel-Rahman WM, Valo S, Mecklin JP, Jarvinen HJ, Gleizes PE, Peltomaki P. Germline Mutation of RPS20, Encoding a Ribosomal Protein, Causes Predisposition to Hereditary Nonpolyposis Colorectal Carcinoma Without

DNA Mismatch Repair Deficiency. Gastroenterology. 2014; 147:595. https://doi.org/10.1053/j.gastro.2014.06.009.

8. Palles C, Cazier JB, Howarth KM, Domingo E, Jones AM, Broderick P, Kemp Z, Spain SL, Almeida EG, Salguero I, Sherborne A, Chubb D, Carvajal-Carmona LG, et al. Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas (vol 45, pg 136, 2013). Nature Genetics. 2013; 45:713. https://doi.org/10.1038/ng0613-713b.

9. Park DJ, Tao K, Le Calvez-Kelm F, Nguyen-Dumont T, Robinot N, Hammet F, Odefrey F, Tsimiklis H, Teo ZL, Thingholm LB, Young EL, Voegele C, Lonie A, et al. Rare mutations in RINT1 predispose carriers to breast and Lynch syndrome-spectrum cancers. Cancer Discov. 2014; 4:804–15. https://doi.org/10.1158/2159-8290.CD-14-0212.

10. Segui N, Mina LB, Lazaro C, Sanz-Pamplona R, Pons T, Navarro M, Bellido F, Lopez-Doriga A, Valdes-Mas R, Pineda M, Guino E, Vidal A, Soto JL, et al. Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. Gastroenterology. 2015; 149:563–6. https://doi.org/10.1053/j.gastro.2015.05.056.

11. Weren RD, Ligtenberg MJ, Kets CM, de Voer RM, Verwiel ET, Spruijt L, van Zelst-Stams WA, Jongmans MC, Gilissen C, Hehir-Kwa JY, Hoischen A, Shendure J, Boyle EA, et al. A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. Nat Genet. 2015; 47:668–71. https://doi.org/10.1038/ng.3287.

12. Gray-McGuire C, Guda K, Adrianto I, Lin CP, Natale L, Potter JD, Newcomb P, Poole EM, Ulrich CM, Lindor N, Goode EL, Fridley BL, Jenkins R, et al. Confirmation of Linkage to and Localization of Familial Colon Cancer Risk Haplotype on Chromosome 9q22. Cancer Research. 2010; 70:5409–18. https://doi.org/10.1158/0008-5472.can-10-0188.

13. Kemp ZE, Carvajal-Carmona LG, Barclay E, Gorman M, Martin L, Wood W, Rowan A, Donohue C, Spain S, Jaeger E, Evans DG, Maher ER, Bishop T, et al. Evidence of linkage to chromosome 9q22.33 in colorectal cancer kindreds from the United Kingdom. Cancer Res. 2006; 66:5003–6. https://doi.org/10.1158/0008-5472.can-05-4074.

14. Wiesner GL, Daley D, Lewis S, Ticknor C, Platzer P, Lutterbaugh J, MacMillen M, Baliner B, Willis J, Elston RC, Markowitz SD. A subset of familial colorectal neoplasia kindreds linked to chromosome 9q22.2-31.2. Proc Natl Acad Sci U S A. 2003; 100:12961–5. https://doi.org/10.1073/pnas.2132286100.

15. Clarke E, Green RC, Green JS, Mahoney K, Parfrey PS, Younghusband HB, Woods MO. Inherited deleterious variants in GALNT12 are associated with CRC susceptibility. Hum Mutat. 2012; 33:1056–8. https://doi.org/10.1002/humu.22088.

16. Guda K, Moinova H, He J, Jamison O, Ravi L, Natale L, Lutterbaugh J, Lawrence E, Lewis S, Willson JK, Lowe JB, Wiesner GL, Parmigiani G, et al. Inactivating germ-line and somatic mutations in polypeptide N-acetylgalactosaminyltransferase 12 in human colon cancers. Proc Natl Acad Sci U S A. 2009; 106:12921–5. https://doi.org/10.1073/pnas.0901454106.

17. Lejeune S, Guillemot F, Triboulet JP, Cattan S, Mouton C, Group P, Porchet N, Manouvrier S, Buisine MP. Low frequency of AXIN2 mutations and high frequency of MUTYH mutations in patients with multiple polyposis. Hum Mutat. 2006; 27:1064. https://doi.org/10.1002/humu.9460.

18. Rivera B, Perea J, Sanchez E, Villapun M, Sanchez-Tome E, Mercadillo F, Robledo M, Benitez J, Urioste M. A novel AXIN2 germline variant associated with attenuated FAP without signs of oligondontia or ectodermal dysplasia. Eur J Hum Genet. 2014; 22:423–6. https://doi.org/10.1038/ejhg.2013.146.

19. Valle L, Serena-Acedo T, Liyanarachchi S, Hampel H, Comeras I, Li Z, Zeng Q, Zhang HT, Pennison MJ, Sadim M, Pasche B, Tanner SM, de la Chapelle A. Germline allele-specific expression of TGFBR1 confers an increased risk of colorectal cancer. Science. 2008; 321:1361–5. https://doi.org/10.1126/science.1159397.

20. Skoglund J, Djureinovic T, Zhou XL, Vandrovcova J, Renkonen E, Iselius L, Bisgaard ML, Peltomaki P, Lindblom A. Linkage analysis in a large Swedish family supports the presence of a susceptibility locus for adenoma and colorectal cancer on chromosome 9q22.32-31.1. J Med Genet. 2006; 43:e7. https://doi.org/10.1136/jmg.2005.033928.

21. Liu W, Jiao X, Thutkawkorapin J, Mahdessian H, Lindblom A. Cancer risk susceptibility loci in a Swedish population. Oncotarget. 2017; 8:110300–110310. https://doi.org/10.18632/oncotarget.22687.

22. Wilming LG, Gilbert JGR, Howe K, Trevanion S, Hubbard T, Harrow JL. The vertebrate genome annotation (Vega) database. Nucleic Acids Res. 2008; 36:D753–D760. https://doi.org/10.1093/nar/gkm987.

23. ENCODE Project Consortium. The ENCODE (ENCyclopedia Of DNA Elements) Project. Science. 2004; 306:636–40. . https://doi.org/10.1126/science.1105136.

24. Gupta RA, Shah N, Wang KC, Kim J, Horlings HM, Wong DJ, Tsai MC, Hung T, Argani P, Rinn JL, Wang Y, Brzoska P, Kong B, et al. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. Nature. 2010; 464:1071–1076. https://doi.org/10.1038/nature08975.

25. Yap KL, Li S, Muñoz-Cabello AM, Raguz S, Zeng L, Mujtaba S, Gil J, Walsh MJ, Zhou MM. Molecular interplay of the noncoding RNA ANRIL and methylated histone H3 lysine 27 by polycomb CBX7 in transcriptional silencing of INK4a. Mol Cell. 2010; 38:662–74. https://doi.org/10.1016/j.molcel.2010.03.021.

26. Kotake Y, Nakagawa T, Kitagawa K, Suzuki S, Liu N, Kitagawa M, Xiong Y. Long non-coding RNA ANRIL is required for the PRC2 recruitment to and silencing of p15(INK4B) tumor suppressor gene. Oncogene. 2011; 30:1956–62. https://doi.org/10.1038/onc.2010.568.

27. Cunnington MS, Santibanez KM, Mayosi BM, Burn J, Keavney B. Chromosome 9p21 SNPs Associated with Multiple Disease Phenotypes Correlate with ANRIL Expression. PLoS Genet. 2010; 6:e1000899. https://doi.org/10.1371/journal.pgen.1000899.

28. Chung JH, Bunz F. A loss-of-function mutation in PTCH1 suggests a role for autocrine hedgehog signaling in colorectal tumorigenesis. Oncotarget. 2014; 4:2208–11. https://doi.org/10.18632/oncotarget.1651.

29. Garcia-Solano J, Garcia-Solano ME, Torres-Moreno D, Carbonell P, Trujillo-Santos J, Perez-Guillermo M, Conesa-Zamora P. Biomarkers for the identification of precursor polyps of colorectal serrated adenocarcinomas. Cell Oncol (Dordr). 2016; 39:243–52. https://doi.org/10.1007/s13402-016-0269-5.

30. Peng L, Hu J, Li S, Wang Z, Xia B, Jiang B, Li B, Zhang Y, Wang J, Wang X. Aberrant methylation of the PTCH1 gene promoter region in aberrant crypt foci. Int J Cancer. 2013; 132:E18–25. https://doi.org/10.1002/ijc.27812.

31. Esteban-Jurado C, Franch-Exposito S, Munoz J, Ocana T, Carballal S, Lopez-Ceron M, Cuatrecasas M, Vila-Casadesus M, Lozano JJ, Serra E, Beltran S, Brea-Fernandez A, Ruiz-Ponte C, et al. The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. Eur J Hum Genet. 2016; 24:1501–5. https://doi.org/10.1038/ejhg.2016.44.

32. Tummala H, Kirwan M, Walne AJ, Hossain U, Jackson N, Pondarre C, Plagnol V, Vulliamy T, Dokal I. ERCC6L2 mutations link a distinct bone-marrow-failure syndrome to DNA repair and mitochondrial function. Am J Hum Genet. 2014; 94:246–56. https://doi.org/10.1016/j.ajhg.2014.01.007.

33. Lagerstedt Robinson K, Liu T, Vandrovcova J, Halvarsson B, Clendenning M, Frebourg T, Papadopoulos N, Kinzler KW, Vogelstein B, Peltomaki P, Kolodner RD, Nilbert M, Lindblom A. Lynch syndrome (hereditary nonpolyposis colorectal cancer) diagnostics. J Natl Cancer Inst. 2007; 99:291–9. https://doi.org/10.1093/jnci/djk051.

34. Magnusson PK, Almqvist C, Rahman I, Ganna A, Viktorin A, Walum H, Halldner L, Lundstrom S, Ullen F, Langstrom N, Larsson H, Nyman A, Gumpert CH, et al. The Swedish Twin Registry: establishment of a biobank and other recent developments. Twin Res Hum Genet. 2013; 16:317–29. https://doi.org/10.1017/thg.2012.104.

35. Barber TD, McManus K, Yuen KW, Reis M, Parmigiani G, Shen D, Barrett I, Nouhi Y, Spencer F, Markowitz S, Velculescu VE, Kinzler KW, Vogelstein B, et al. Chromatid cohesion defects may underlie chromosome instability in human colorectal cancers. Proc Natl Acad Sci U S A. 2008; 105:3443–8. https://doi.org/10.1073/pnas.0712384105.

36. Sjöblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber TD, Mandelker D, Leary RJ, Ptak J, Silliman N, Szabo S, Buckhaults P, Farrell C, et al. The consensus coding sequences of human breast and colorectal cancers. Science. 2007; 314:268–74. https://doi.org/10.1126/science.1133427.

37. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009; 25:1754–60. https://doi.org/10.1093/bioinformatics/btp324.

38. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet. 2011; 43:491–8. https://doi.org/10.1038/ng.806.

39. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010; 38:e164. https://doi.org/10.1093/nar/gkq603.

40. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. dbSNP: the NCBI database of genetic variation. Nucleic Acids Research. 2001; 29:308–11. https://doi.org/10.1093/nar/29.1.308.

41. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015; 526:68–74. https://doi.org/10.1038/nature15393.

42. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, Sham PC. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007; 81:559–75. https://doi.org/10.1086/519795.

43. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20:1297–303. https://doi.org/10.1101/gr.107524.110.

44. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J, Banks E, Garimella KV, Altshuler D, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics. 2013; 43:11 0 1–33. https://doi.org/10.1002/0471250953.bi1110s43.