Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Cancer Epidemiol. Author manuscript; available in PMC 2022 June 01.

Published in final edited form as:

Cancer Epidemiol. 2021 June ; 72: 101941. doi:10.1016/j.canep.2021.101941.

An intronic variant in the *CELF4* gene is associated with risk for colorectal cancer

Craig C. Teerlink[a], **Jeff Stevens**[a], **Rolando Hernandez**[b], **Julio C. Facelli**[b,c], **Lisa A. Cannon-Albright**[a,d,e]

[a.]Department of Internal Medicine, University of Utah School of Medicine, Salt Lake City, UT 84132, USA.

[b.]Department of Biomedical Informatics, University of Utah School of Medicine, Salt Lake City, UT 84108, USA.

[c.]Center for Clinical and Translational Science, University of Utah School of Medicine, Salt Lake City, UT 84108, USA

[d.]George E. Wahlen Department of Veterans Affairs Medical Center, Salt Lake City, UT 84148, USA

[e.]Huntsman Cancer Institute, Salt Lake City, UT 84112, USA

Abstract

Background: Germline predisposition variants associated with colorectal cancer (CRC) have been identified but all are not yet identified. We sought to identify the responsible predisposition germline variant in an extended high-risk CRC pedigree that exhibited evidence of linkage to the 18q12.2 region (TLOD = +2.81).

Methods: DNA from two distantly related carriers of the hypothesized predisposition haplotype on 18q12.2 was sequenced to identify candidate variants. The candidate rare variants shared by the related sequenced subjects were screened in 3,094 CRC cases and 5x population-matched controls from UKBiobank to test for association. Further segregation of the variant was tested via Taqman assay in other sampled individuals in the pedigree.

Corresponding author Craig C. Teerlink, PhD, 295 Chipeta Way, Salt Lake City, UT 84108; Phone: (801)587-9303; craig.teerlink@utah.edu.

Author Contributions

Craig Teerlink: Methodology, Formal analysis, Investigation, Data curation, Writing – Original draft, Writing – Review and Editing. **Jeff Stevens**: Formal analysis, Investigation, Data curation, Writing – Original draft, Writing – Review and Editing. **Rolando Hernandez**: Formal Analysis, Investigation, Writing - Original Draft, Writing - Review and Editing, Visualization. **Julio Facelli**: Conceptualization, Software, Resources Writing-Review Editing. **Lisa Cannon-Albright**: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – Original draft, Writing - Review and editing, Visualization, Supervision, Project administration.

Declarations of interest: none.

DATA STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

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 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.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results: Analysis of whole genome sequence data for the two related hypothesized predisposition haplotype carriers, restricted to the shared haplotype boundaries, identified multiple (n=6) rare candidate non-coding variants that were tested for association with CRC risk in UKBiobank. A rare intronic variant of *CELF4* gene, rs568643870, was significantly associated with CRC (p=0.004, OR=5.0), and segregated with CRC in other members of the linked pedigree.

Conclusion: Evidence of segregation in a high-risk pedigree, case-control association in an external dataset, and identification of additional CRC-affected carriers in the linked pedigree support a role for a rare *CELF4* intronic variant in CRC risk.

Keywords

Linkage analysis; colorectal cancer; *CELF4*; UPDB; high-risk pedigree

1. INTRODUCTION

High-risk pedigree studies are a powerful method for identification of rare predisposition variants with large effects; although because pedigree resources are rare and difficult to acquire, genetic association studies of large numbers of cases and controls have become a more prevalent approach. It remains true that pedigree studies should be performed when these rare resources are available [1,2], and that pedigree studies and association studies can complement each other in the search for disease predisposition genes and variants. Unique resources in Utah, including a genealogy of the majority of the state that dates back centuries and is linked to decades of statewide cancer records for the state, have allowed the identification and study of large numbers of high-risk cancer pedigrees that are informative for predisposition gene identification [3–7]. Here we pursued evidence of linkage in a high-risk colorectal cancer (CRC) pedigree with focused whole genome sequencing to identify a strong candidate variant for CRC risk.

2. MATERIALS AND METHODS

2.1 Ethical considerations

For Utah subjects, the University of Utah Institutional Review Board approved these studies and informed consent was obtained for all subjects. For UKBiobank subjects, all data was de-identified and individual subject's permissions were not necessary.

2.2 High-risk CRC pedigrees from the Utah Population Data Base (UPDB)

The UPDB includes a computerized genealogy of Utah from its 19th century founders to modern day. It has been linked to the statewide Utah Cancer Registry from 1966, which has been an NCI Surveillance, Epidemiology, and End-Results (SEER) Registry from 1973. This has allowed identification of extended Utah pedigrees exhibiting significantly increased risk for CRC. Over 4,000 CRC cases and relatives in almost 300 high-risk CRC pedigrees in UPDB have been recruited, consented, and sampled over several decades.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

2.3 Genotypes

Genome-wide genotype data for ~700,000 markers (Illumina OmniExpress platform) was available for 106 CRC cases belonging to 24 high-risk CRC pedigrees. Genome wide linkage analysis performed on these pedigrees identified regions of interest for CRC predisposition. DNA samples from an additional 65 pedigree members in these 24 pedigrees (without dense SNP genotyping data) were available for segregation testing of candidate variants. Status of all CRC cases was confirmed in the Utah SEER Cancer Registry.

Quality control of genetic markers for linkage analysis using PLINK software [8] included removal of subjects with call rate < 98% and removal of markers with call rate < 98%, minor allele frequency (MAF) < 1%, or with significant deviation from Hardy-Weinberg equilibrium (HWE) ($p < 1e^{-5}$). Markers were then reduced to a set of ~27K markers not in linkage disequilibrium ($r^2 < 0.1$ with another marker) and with heterozygosity > 0.3, which has been previously shown to produce an independent set of markers suitable for linkage analysis [9].

2.4 Linkage analysis

Linkage analysis used a robust multipoint linkage statistic referred to as the TLOD and implemented in MCSIM software [10]. The TLOD statistic uses multipoint information but also optimizes the LOD score over the recombination fraction, similar to a two-point LOD score, allowing for robustness to model misspecification [11]. Pedigrees were analyzed using both a general parametric dominant and recessive model. Given their extended nature, pedigrees were assumed to be singly informative for linkage and were analyzed individually. Statistical significance was interpreted according to established guidelines with LOD > 3.3 considered significant and LOD > 1.89 suggestive [12].

2.5 Genome sequencing

Two carriers of the hypothesized predisposition haplotype in the pedigree with suggestive evidence of linkage at 18q12.2 were genome sequenced (GS) to identify rare genetic variants occurring on the shared haplotype. GS data was generated by NantOmics LLC. A DNA library was prepared from 2 micrograms of DNA per sample using the Illumina TruSeq DNA PCR-Free GS library kit. Samples were run on the Illumina HiSeq 2000 instrument that generates paired end reads of up to 150 base pairs in length to an average read depth of 58x coverage. FastQ files were processed by the Utah Genome project and mapped to the human genome GRCh37 reference genome using BWA MEM [13]. Variants were called using Genome Analysis Toolkit 3.5.0 [14] software following Broad Institute Best Practices Guidelines. Variants were annotated with Annovar software [15]. Candidate variants were filtered on the criteria of being rare in the population (MAF<0.005), observed in both sequenced subjects, and located within the linked region.

2.6 UKBiobank CRC cases and controls

Candidate variants were analyzed for CRC risk association in a set of 3,094 Caucasian CRC cases and 15,470 ancestrally matched controls from the UKBiobank's 488,377 total subjects genotyped on the Illumina OmniExpress SNP array [16]. UKBiobank case and control subjects were matched via principal components (PCs) using ~27K independent markers

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

that excluded several genomic regions known to adversely affect PC analysis [17]. FLASHPCA2 software was used to generate eigenvectors for control selection [17]. Controls were selected from among 191,466 Caucasian UKBiobank subjects who were over age 70 years of age and had no cancer diagnosis. Five controls, representing the nearest neighbors based on Euclidean distance of the first two PCs, were selected for each case. Supplemental Figure 1 shows UKBiobank cases and controls plotted on first PC versus second PC.

2.7 UKBiobank Imputation

The selected UKBiobank case and control subjects were imputed to ~40M SNP markers using the Haplotype Reference Consortium's (HRC) 67K background genomes [18]. Beginning with 784,256 observed SNP genotypes, pre-imputation quality control using PLINK software [8] required sample genotyping >98% (no subjects removed). A total of 353,578 markers were removed by filtering for genotyping call rate <98%, HWE $p<1e^{-5}$, MAF<0.005, duplicated position in the HRC's reference genome, or site not included in the HRC's reference genome. The remaining 430,678 SNPs were converted to human genome B37 forward strand orientation using GenotypeHarmonizer software [19] and served as the basis for imputation. Imputation was performed with EAGLE v2.3 software for phasing [20] and MINIMAC3 software for imputation [21]. Post-imputation quality control included removing markers with imputation information score (INFO-r^2) <0.7 [22–24].

2.8 Segregation testing

The *CELF4* variant that emerged from GS data in the linked region and was significantly associated with CRC among the UKBiobank CRC case and control subjects was confirmed in the original linked subjects and evaluated for segregation to additional sampled relatives in the linked pedigree using a custom Taqman assay. We used the RVsharing probability [25] to express the probability of the observed configuration of carriers and affection status in the pedigree having occurred by chance, assuming the variant is rare (MAF <1%) and entered the pedigree only once.

2.9 RNA structure prediction

The non-coding, intronic *CELF4* variant (dbSNP id: rs568643870, position chr18:34932587, HGVS id: NC_000018.9:g.34932587T>A; NC_000018.10:g.37352624T>A) in *CELF4* was examined using RNA structure prediction. The genomic reference sequence used for the analysis (NC_000018.10 Reference GRCh38.p13 Primary Assembly) was retrieved from NCBI and modified in UGENE to prepare the wild type and variant sequences [26]. This was done by selecting 100 base pairs to either side of the affected position, along with swapping T -> U to simulate the local section of the precursor mRNA before the splicing process. It is noteworthy that in the wild type, the nucleotides at the position are UU**U**AA, which is changed to U**UAA**A (bolded nucleotides form an ochre stop codon) in the variant. Depending on the reading frame, this could be highly deleterious and a potential cause of pathogenesis. The *CELF4* wildtype and variant sequences were used for RNA 3D structure prediction using the RNAComposer server with RNAFold secondary structure prediction [27]. The two resulting structures were compared using UCSF Chimera [28].

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

3. RESULTS

One pedigree in the genome-wide CRC linkage scan showed significant evidence of linkage to chromosome 22 and was discussed elsewhere [29]. The pedigree with the second highest LOD score in the linkage scan achieved suggestive evidence for linkage with a TLOD score of +2.81 (RVsharing p-value=$1.6e^{-4}$) on chromosome 18q12.2 under the dominant model with a 1-LOD support interval of 22.0–34.9 Mb (hg19).

The female founder of the linked pedigree (Figure 1) was born in Kentucky in the early 19th century and has almost 10,000 descendants recorded in the UPDB. These descendants include 37 CRC cases (18.3 expected; p=$8e^{-5}$); the only other cancer observed in excess among these descendants is prostate cancer (n=54 observed, 37.3 expected; p=$6e^{-3}$).

No coding variants and 89 non-coding variants were identified as shared by the two sequenced subjects in the linked region and were selected for case/control risk association testing in the UKBiobank CRC case and control subjects. Only six of the 89 rare variants appeared among the imputed UKBiobank genomes and passed the Info-r^2 data quality filter of 0.7 (Bonferonni corrected p-value threshold of 0.008). After correction, one variant, rs568643870, intronic to the *CELF4* gene showed significant evidence for association with CRC risk in the UKBiobank cases and controls, with OR= 5.0 (Fisher's exact p = 0.004). The *CELF4* variant (rs568643870) has a population MAF of 0.001 (31/31,038 genomes) in genomAD across all ethnic groups [30], and a GERP score of −0.01 indicating low conservation across species [31]. Variant rs568643870 was then tested via Taqman assay in 14 sampled relatives from the linked pedigree, including the originally genotyped individuals. The original hypothesized predisposition haplotype carriers from the linkage analysis were confirmed to carry the variant, as expected, and 2 additional sampled carriers were identified: one additional CRC case carrier and the 1 carrier daughter of another CRC case (RVsharing p-value = $6.5e^{-6}$). The chromosome 18q-linked pedigree depicted in Figure 1 indicates the carriers of the hypothesized risk haplotype who formed the basis of the linkage results, the 2 haplotype carriers from the pedigree on whom GS was performed, and all confirmed variant carriers, including 2 additional sampled affected carriers identified by Taqman assay in the pedigree.

Results of the RNA structure analysis are depicted in Figures 2 through 5. The tan structure with the green highlight is the predicted structure for the wild type sequence in Figure 2 and the blue structure with red patch in Figure 3 is the predicted variant sequence. The patches designate the nucleic acid position that changes in the variant sequence. In the images provided, the two structures resemble each other very closely with only the residues around the variant position undergoing large conformational changes (Figures 4 and 5). When superimposed, there is an RMSD of 19.03 Å across all pairs between the structures, due to the large structural change in the loop region. According to the final RNAComposer results after structural refinement, there is an approximately −49.6 (kcal/mol) difference in total energy between the wild type (−4025.224 kcal/mol) and variant (−4074.821 kcal/mol) structures; based on this, it is highly likely that the variant is destabilizing to the RNA structure.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

4. DISCUSSION

While GWAS analysis is commonly used to identify new colorectal cancer predisposition variants, it is not powerful for identification of rare variants. Extended high-risk CRC pedigrees, while rare, are informative for such identification. Studies of extended Utah high-risk pedigrees have identified multiple predisposition genes and variants for cancer and other phenotypes [3–7, 32–38]. Here we pursued suggestive evidence of linkage from a genome-wide scan that included a small set of extended high-risk CRC pedigrees with available dense genotype data. Although linkage evidence for the CRC pedigree was only suggestive, such evidence is rare for extended pedigrees, and was therefore pursued with GS analysis of 2 distantly related individuals hypothesized to share the predisposition variant giving rise to evidence of linkage. Analysis of the GS data in these 2 individuals identified a small number of candidate predisposition variants; a subset of which had population level genotypes available (e.g. UKBiobank data) for testing association of each variant with CRC risk. Only one of the candidate variants showed significant association with CRC risk after multiple testing correction, *CELF4* rs568643870. The variant was confirmed in the original CRC cases analyzed for linkage, as expected, and was also identified in an additional CRC case and the unaffected daughter of another CRC case who was inferred to carry.

RNA structure analysis predicted total energy differences that highly destabilize RNA structure of fragments containing the variant. Furthermore, due to the destabilizing nature of the variant, it is possible that this abnormal loop and energetic change in the variant structure could result in abnormal splicing of the pre-mRNA, which could then introduce an in-frame stop codon into the spliced transcript. This would result in detection by the nonsense-mediated decay machinery in the cytoplasm and the transcript would be eliminated, so the CELF4 protein would not be expressed. The combined evidence for this *CELF4* variant strongly suggests it is associated with increased risk for CRC.

CELF4 is ubiquitously expressed with high expression in the brain [39], however, the variant rs568643870 has not been evaluated as an expression quantitative trait locus in GETx. Histone modification by ChIP-seq experiments have shown a GM12878 signal at this variant suggesting the wild type of this variant may interact with two RNA binding proteins PABPC1 and ELAV1 [40]. RNA binding proteins (RBPs) bind to single or double stranded RNA following the RNA transcription process. RBPs contain sequence specific structural motifs that can be involved in numerous functions such as alternate splicing, RNA editing, export, mRNA localization, and translation [41]. The RNA binding protein PABPC1 has been associated with several cancers including CRC [42–44]. The RNA binding protein ELAV1 has also been associated with CRC [44]. The RNA prediction modeling results indicate that, due to the destabilization of the variant RNA sequences, it is likely that RNA binding to ELAV1 and PABPC1 may be compromised. Furthermore, in 109 CRC cases, DNA copy number loss of a 1-Mb segment of 18q12.2 containing only the *CELF4* gene was significantly correlated with aggressive growth of CRC, nominating copy number loss of *CELF4* as a prognostic indicator of CRC [45].

This analysis identifying a CRC risk associated variant in *CELF4* has strengths and limitations. Strengths include the high-quality cancer data in the Utah population combined

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

with extended genealogy data, resulting in powerful genetic resources for predisposition gene identification which reduce potential bias from recall and ascertainment. Validation of association with this risk variant with CRC risk in the independent UKBiobank population, as well as confirmation of segregation of the variant in additional CRC cases in the pedigree in which it was observed, both strongly enhance the original suggestive evidence of linkage. The large energy de-equilibration observed in RNA fragments in the region of the variant indicates that the nucleotide structure of this region may be very sensitive to relatively small changes in the sequence affecting proper translation of the associated exons. Weaknesses of the study include the small number of high-risk CRC pedigrees analyzed, that linkage evidence was only suggestive, a general lack of information regarding the function of the *CELF4* gene, and the possibility that the variant identified is not causal but rather is in linkage disequilibrium with some unobserved proximal marker that shows stronger biological relevance. Finally, because the modeling study included arbitrary selection of possible RNA fragments including the variant, these fragments may not be representative of the actual experimental RNA fragments that are determined by the splicing of the gene.

5. CONCLUSIONS

In addition to identifying strong evidence for a CRC predisposition variant in *CELF4*, this study reinforces the value of high-risk pedigree studies for rare risk predisposition variant identification and validation, and supports similar analyses of extended high-risk cancer pedigrees in Utah and elsewhere. The complementary use of in silico methods highlight the importance of using these emerging techniques to provide further support to genetic epidemiology findings.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge donation of whole genome sequencing by the Utah Genome Project for whole genome sequencing.

FUNDING SOURCES

This work was supported by the Utah Cancer Registry, which is funded by the National Cancer Institute's SEER Program, Contract No. HHSN261201800016I, the US Center for Disease Control and Prevention's National Program of Cancer Registries, Cooperative Agreement No. NU58DP0063200-01, with additional support from the University of Utah and Huntsman Cancer Foundation. Partial support for all datasets within the Utah Population Database is provided by the University of Utah, Huntsman Cancer Institute and the Huntsman Cancer Institute Cancer Center Support grant, P30 CA42014 from the National Cancer Institute. Additional support from the National Institutes of Health included T15LM00712418 to RH, 1S10OD02164401A1 to the Utah Center for High Performance Computing that provided computational resources, and 1ULTR002538 to JCF. LACA was partially supported by the Huntsman Cancer Institute Cancer Center Support grant, P30 CA42014 from the National Cancer Institute. Funding sources were not involved in the design of the study, collection, analysis and interpretation of data, in the writing of the report, or in the decision to submit the article for publication.

REFERENCES

1. Manolio TA, Collins FA, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, & Visscher PM (2009). Finding the missing heritability of complex diseases. Nature, 461, 747–753. *doi:* 10.1038/nature08494 [PubMed: 19812666]

2. Wijsman EM (2012). The role of large pedigrees in an era of high-throughput sequencing. Hum Genet, 131, 1555–1563. *doi:* 10.1007/s00439-012-1190-2 [PubMed: 22714655]
3. Miki Y, Swensen J, Shattuck-Eidens D, Futreal PA, Harshman K, Tavtigian S, Liu Q, Cochran C, Bennett LM, & Ding W (1994). A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science, 266, 66–71. *doi:* 10.1126/science.7545954 [PubMed: 7545954]
4. Tavtigian SV, Simard J, Rommens J, Couch F, Shattuck-Eidens D, Neuhausen S, Merajver S, Thorlacius S, Offit K, Stoppa-Lyonnet D, Belanger C, Bell R, Berry S, Bogden R, Chen Q, Davis T, Dumont M, Frye C, Hattier T, Jammulapati S, Janecki T, Jiang P, Kehrer R, Leblanc JF, Mitchell JT, McArthur-Morrison J, Nguyen K, Peng Y, Samson C, Schroeder M, Snyder SC, Steele L, Stringfellow M, Stroup C, Swedlund B, Swense J, Teng D, Thomas A, Tran T, Tranchant M, Weaver-Feldhaus J, Wong AK, Shizuya H, Eyfjord JE, Cannon-Albright L, Tranchant M, Labrie F, Skolnick MH, Weber B, Kamb A, & Goldgar DE (1996). The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. Nat Genet, 12, 333–337. *doi:* 10.1038/ng0396-333 [PubMed: 8589730]
5. Tavtigian SV, Simard J, Teng DH, Abtin V, Baumgard M, Beck A, Camp NJ, Carillo AR, Chen Y, Dayananth P, Desrochers M, Dumont M, Farnham JM, Frank D, Frye C, Ghaffari S, Gupte JS, Hu R, Iliev D, Janecki T, Kort EN, Laity KE, Leavitt A, Leblanc G, McArthur-Morrison J, Pederson A, Penn B, Peterson KT, Reid JE, Richards S, Schroeder M, Smith R, Snyder SC, Swedlund B, Swensen J, Thomas A, Tranchant M, Woodland AM, Labrie F, Skolnick MH, Neuhausen S, Rommens J, & Cannon-Albright LA. (2001). A candidate prostate cancer predisposition gene at 17p. Nat Genet, 27, 172–180. *doi:* 10.1038/84808 [PubMed: 11175785]
6. Kamb A, Shattuck-Eidens D, Eeles R, Liu Q, Gruis NA, Ding W, Hussey C, Tran T, Miki Y, Weaver-Feldhaus J, McClure M, Aitken JF, Anderson DE, Bergman W, Frants R, Goldgar DE, Green A, MacLennan R, Martin NG, Meyer LJ, Youl P, Zone JJ, Skolnick MH, & Cannon-Albright LA (1994). Analysis of the p16(CDKN2) as a candidate for the chromosome 9p melanoma susceptibility locus. Nat Genet, 8, 22–26. *doi:* 10.1038/ng0994-22
7. Teerlink CC, Huff C, Stevens J, Yu Y, Holmen SL, Silvis MR, Trombetti K, Zhao H, Grossman D, Farnham JM, Wen J, Facelli JC, Thomas A, Babst M, Florell SR, Meyer L, Zone JJ, Leachman S, & Cannon-Albright LA (2018). A nonsynonymous variant in the GOLM1 gene in cutaneous malignant melanoma. J Natl Cancer Inst, 110, 1380–1385. *doi:* 10.1093/jnci/djy058 [PubMed: 29659923]
8. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, & Sham. P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet, 81, 559–575. doi: 10.1086/519795 [PubMed: 17701901]
9. Allen-Brady K, Horne BD, Malhotra A, Teerlink C, Camp NJ, & Thomas A (2007). Analysis of high-density single-nucleotide polymorphism data: three novel methods that control for linkage disequilibrium between markers in a linkage analysis. BMC Proc, 1 Suppl, 1:S160. *doi:* 10.1186/1753-6561-1-s1-s160 [PubMed: 18466506]
10. Abkevich V, Camp NJ, Gutin A, Farnham JM, Cannon-Albright L, & Thomas A (2001). A robust multipoint linkage statistic (TLOD) for mapping complex trait loci. Genet Epidemiol, 21, Suppl 1, S492–S497. *doi:* 10.1002/gepi.2001.21.s1.s492 [PubMed: 11793725]
11. Göring HH, & Terwilliger JD (2000). Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. Am J Hum Genet, 66, 1298–1309. *doi:* 10.1086/302846 [PubMed: 10731467]
12. Lander E, & Kruglyak L (1995). Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet, 11, 241–247. *doi:* 10.1038/ng1195-241 [PubMed: 7581446]
13. Li H, & Durbin R (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754–1760. *doi:* 10.1093/bioinformatics/btp324 [PubMed: 19451168]
14. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, & DePristo MA (2010). The Genome Analysis Toolkit: a MapReduce

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

framework for analyzing next-generation DNA sequencing data. Genome Res, 20, 1297–1303. *doi:* 10.1101/gr.107524.110 [PubMed: 20644199]

15. Wang K, Li M, & Hakonarson H (2010). ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Research, 38, e164. *doi:* 10.1093/nar/gkq603 [PubMed: 20601685]
16. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, Downey P, Elliott P, Green J, Landray M, Liu B, Matthews P, Ong G, Pell J, Silman A, Young A, Sprosen T, Peakman T, & Collins R (2015). UK Biobank: An open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med, 12, e1001779. *doi:* 10.1371/journal.pmed.1001779 [PubMed: 25826379]
17. Abraham G, Qiu Y, & Inouye M (2017). FlashPCA2: principal component analysis of Biobank-scale genotype datasets. Bioinformatics, 33, 2776–2778. *doi:* 10.1093/bioinformatics/btx299 [PubMed: 28475694]
18. The Haplotype Reference Consortium. (2016). A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet, 48, 1279–1283. *doi:* 10.1038/ng.3643 [PubMed: 27548312]
19. Deelen P, Bonder MJ, van der Velde KJ, Westra H, Winder E, Hendriksen D, Franke L, & Swertz MA (2014). Genotype harmonizer: automatic strand alignment and format conversion for genotype data integration. BMC Research Notes, 7, 901. *doi:* 10.1186/1756-0500-7-901 [PubMed: 25495213]
20. Loh PR, Palamara PF, & Price AL (2016). Fast and accurate long-range phasing in a UK Biobank cohort. Nat Genet, 48, 811–816. *doi:* 10.1038/ng.3571 [PubMed: 27270109]
21. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, Vrieze SI, Chew EY, Levy S, McGue M,Schlessinger D, Stambolian D, Loh P, Iacono WG, Swaroop A, Scott LJ, Cucca F, Kronenberg F, Boehnke M, Abecasis GR, & Fuchsberger C (2016). Next-generation genotype imputation service and methods. Nat Genet, 48, 1284–1287. *doi:* 10.1038/ng.3656 [PubMed: 27571263]
22. Ziv E, Dean E, Hu D, Martino A, Serie D, Curtin K, Campa D, Aftab B, Bracci P, Buda G, Zhao Y, Caswell-Jin J, Diasio R, Dumontet C, Dudziński M, Fejerman L, Greenberg A, Huntsman S, Jamroziak K, Jurczyszyn A, Kumar S, Atanackovic D, Glenn M, Cannon-Albright LA, Jones B, Lee A, Marques H, Martin T, Martinez-Lopez J, Rajkumar V, Sainz J, Vangsted AJ, Wątek M, Wolf J, Slager S, Camp NJ, Canzian F, & Vachon C (2015). Genome-wide association study identifies variants at 16p13 associated with survival in multiple myeloma patients. Nat Communications, 6, 7539. *doi:* 10.1038/ncomms8539
23. Schumacher FR, Al Olama AA, Berndt SI, Benlloch S, Ahmed M, Saunders EJ, Dadaev T, Leongamornlert D, Anokian E, Cieza-Borrella C, Goh C, Brook MN, Sheng X, Fachal L, Dennis J, Tyrer J, Muir K, Lophatananon A, Stevens VL, Gapstur SM, Carter BD, Tangen CM, Goodman PJ, Thompson IM, Batra J, Chambers S, Moya L, Clements J, Horvath L, Tilley W, Risbridger GP, Gronberg H, Aly M, Nordström T, Pharoah P, Pashayan N, Schleutker J, Tammela TLJ, Sipeky C, Auvinen A, Albanes D, Weinstein S, Wolk A, Håkansson N, West CML, Dunning AM, Burnet N, Mucci LA, Giovannucci E, Andriole GL, Cussenot O, Cancel-Tassin G, Koutros S, Beane Freeman LE, Sorensen KD, Orntoft TF, Borre M, Maehle L, Grindedal EM, Neal DE, Donovan JL, Hamdy FC, Martin RM, Travis RC, Key TJ, Hamilton RJ, Fleshner NE, Finelli A, Ingles SA, Stern MC, Rosenstein BS, Kerns SL, Ostrer H, Lu YJ, Zhang HW, Feng N, Mao X, Guo X, Wang G, Sun Z, Giles GG, Southey MC, MacInnis RJ, FitzGerald LM, Kibel AS, Drake BF, Vega A, Gómez-Caamaño A, Szulkin R, Eklund M, Kogevinas M, Llorca J, Castaño-Vinyals G, Penney KL, Stampfer M, Park JY, Sellers TA, Lin HY, Stanford JL, Cybulski C, Wokolorczyk D, Lubinski J, Ostrander EA, Geybels MS, Nordestgaard BG, Nielsen SF, Weischer M, Bisbjerg R, Røder MA, Iversen P, Brenner H, Cuk K, Holleczek B, Maier C, Luedeke M, Schnoeller T, Kim J, Logothetis CJ, John EM, Teixeira MR, Paulo P, Cardoso M, Neuhausen SL, Steele L, Ding YC, De Ruyck K, De Meerleer G, Ost P, Razack A, Lim J, Teo SH, Lin DW, Newcomb LF, Lessel D, Gamulin M, Kulis T, Kaneva R, Usmani N, Singhal S, Slavov C, Mitev V, Parliament M, Claessens F, Joniau S, Van den Broeck T, Larkin S, Townsend PA, Aukim-Hastie C, Gago-Dominguez M, Castelao JE, Martinez ME, Roobol MJ, Jenster G, van Schaik RHN, Menegaux F, Truong T, Koudou YA, Xu J, Khaw KT, Cannon-Albright L, Pandha H, Michael A, Thibodeau SN, McDonnell SK, Schaid DJ, Lindstrom S, Turman C, Ma J, Hunter DJ, Riboli E, Siddiq A,, Canzian F, Kolonel LN, Le Marchand L, Hoover RN, Machiela MJ, Cui Z, Kraft P, Amos CI, Conti DV, Easton DF, Wiklund F, Chanock SJ, Henderson BE, Kote-Jarai Z, Haiman CA, Eeles RA; Profile Study; Australian

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Prostate Cancer BioResource (APCB); IMPACT Study; Canary PASS Investigators; Breast and Prostate Cancer Cohort Consortium (BPC3); PRACTICAL (Prostate Cancer Association Group to Investigate Cancer-Associated Alterations in the Genome) Consortium; Cancer of the Prostate in Sweden (CAPS); Prostate Cancer Genome-wide Association Study of Uncommon Susceptibility Loci (PEGASUS); & Genetic Associations and Mechanisms in Oncology (GAME-ON)/ Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE) Consortium. (2018). Association analyses of more than 140,000 men identify 63 new prostate cancer susceptibility loci. Nat Genet, 50, 928–936. *doi:* 10.1038/s41588-018-0142-8 [PubMed: 29892016]

24. Huyghe JR, Blen SA, Harrison TA, Kang HM, Chen S, Schmit SL, Conti DV, Qu C, Jeon J, Edlund CK, Greenside P, Wainberg M, Schumacher FR, Smith JD, Levine DM, Nelson SC, Sinnott-Armstrong NA, Albanes D, Alonso MH, Anderson K, Arnau-Collell C, Arndt V, Bamia C, Banbury BL, Baron JA, Berndt SI, Bézieau S, Bishop DT, Boehm J, Boeing H, Brenner H, Brezina S, Buch S, Buchanan DD, Burnett-Hartman A, Butterbach K, Caan BJ, Campbell PT, Carlson CS, Castellví-Bel S, Chan AT, Chang-Claude J, Chanock SJ, Chirlaque MD, Cho SH, Connolly CM, Cross AJ, Cuk K, Curtis KR, de la Chapelle A, Doheny KF, Duggan D, Easton DF, Elias SG, Elliott F, English DR, Feskens EJM, Figueiredo JC, Fischer R, FitzGerald LM, Forman D, Gala M, Gallinger S, Gauderman WJ, Giles GG, Gillanders E, Gong J, Goodman PJ, Grady WM, Grove JS, Gsur A, Gunter MJ, Haile RW, Hampe J, Hampel H, Harlid S, Hayes RB, Hofer P, Hoffmeister M, Hopper JL, Hsu WL, Huang WY, Hudson TJ, Hunter DJ, Ibañez-Sanz G, Idos GE, Ingersoll R, Jackson RD, Jacobs EJ, Jenkins MA, Joshi AD, Joshu CE, Keku TO, Key TJ, Kim HR, Kobayashi E, Kolonel LN, Kooperberg C, Kühn T, Küry S, Kweon SS, Larsson SC, Laurie CA, Le Marchand L, Leal SM, Lee SC, Lejbkowicz F, Lemire M, Li CI, Li L, Lieb W, Lin Y, Lindblom A, Lindor NM, Ling H, Louie TL, Männistö S, Markowitz SD, Martín V, Masala G, McNeil CE, Melas M, Milne RL, Moreno L, Murphy N, Myte R, Naccarati A, Newcomb PA, Offit K, Ogino S, Onland-Moret NC, Pardini B, Parfrey PS, Pearlman R, Perduca V, Pharoah PDP, Pinchev M, Platz EA, Prentice RL, Pugh E, Raskin L, Rennert G, Rennert HS, Riboli E, Rodríguez-Barranco M, Romm J, Sakoda LC, Schafmayer C, Schoen RE, Seminara D, Shah M, Shelford T, Shin MH, Shulman K, Sieri S, Slattery ML, Southey MC, Stadler ZK, Stegmaier C, Su YR, Tangen CM, Thibodeau SN, Thomas DC, Thomas SS, Toland AE, Trichopoulou A, Ulrich CM, Van Den Berg DJ, van Duijnhoven FJB, Van Guelpen B, van Kranen H, Vijai J, Visvanathan K, Vodicka P, Vodickova L, Vymetalkova V, Weigl K, Weinstein SJ, White E, Win AK, Wolf CR, Wolk A, Woods MO, Wu AH, Zaidi SH, Zanke BW, Zhang Q, Zheng W, Scacheri PC, Potter JD, Bassik MC, Kundaje A, Casey G, Moreno V, Abecasis GR, Nickerson DA, Gruber SB, Hsu L, & Peters U (2019). Discovery of common and rare genetic risk variants for colorectal cancer. Nat Genet, 51, 76–87. *doi:* 10.1038/s41588-018-0286-6 [PubMed: 30510241]
25. Bureau A, Younkin SG, Parker MM, Bailey-Wilson JE, Marazita ML, Murray JC, Mangold E, Albacha-Hejazi H, Beaty TH, & Ruczinski I (2014). Inferring rare disease risk variants based on exact probabilities of sharing by multiple affected relatives. Bioinformatics, 30, 2189–2196. *doi:* 10.1093/bioinformatics/btu198 [PubMed: 24740360]
26. Okonechnikov K, Golosova O, Fursov M, Ugene Team. (2012). Unipro UGENE: a unified bioinformatics toolkit. Bioinformatics, 28, 1166–1167. *doi:* 10.1093/bioinformatics/bts091 [PubMed: 22368248]
27. Biesiada M, Purzycka KJ, Szachniuk M, Blazewicz J, & Adamiak RW (2016). Automated RNA 3D structure prediction with RNAComposer. In: Turner DH, Matthews DH, editors. RNA Structure Determination. New York: Springer, p.195–215.
28. Petterson EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, & Ferrin TE (2004). UCSF Chimera – a visualization system for exploratory research and analysis. J Comp Chem, 25, 1605–1612. doi: 10.1002/jcc.20084 [PubMed: 15264254]
29. Teerlink C, Nelson Q, Burt R, & Cannon-Albright L (2014). Significant evidence of linkage for a gene predisposing to colorectal cancer and multiple primary cancers on 22q11. Clin Transl Gastroenterol, 5, e50. *doi:* 10.1038/ctg.2014.1 [PubMed: 24572700]
30. Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, The Genome Aggregation Database Consortium, Neale BM, Daly MJ, & MacArthur DG (2019). Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. bioRXiv, *doi:* 10.1101/531210

31. Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, & Batzoglou S (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. PLoS Comput Biol, 6, e1001025. *doi:* 10.1371/journal.pcbi.1001025 [PubMed: 21152010]
32. Ridge PG, Karch CM, Hsu S, Arano I, Teerlink CC, Ebbert MTW, Gonzalez Murcia JD, Farnham JM, Damato AR, Allen M, Wang X, Harari O, Fernandez VM, Guerreiro R, Bras J, Hardy J, Munger R, Norton M, Sassi C, Singleton A, Younkin SG, Dickson DW, Golde TE, Price ND, Ertekin-Taner N, Cruchaga C, Goate AM, Corcoran C, Tschanz J, Cannon-Albright LA, & Kauwe JSK (2017). Linkage, whole genome sequence, and biological data implicate variants in RAB10 in Alzheimer's disease resilience. Genome Med, 9, 100. *doi:* 10.1186/s13073-017-0486-1 [PubMed: 29183403]
33. Patel D, Mez J, Vardarajan BN, Staley L, Chung J, Zhang X, Farrell JJ, Rynkiewicz MJ, Cannon-Albright LA, Teerlink CC, Stevens J, Corcoran C, Gonzalez Murcia JD, Lopez OL, Mayeux R, Haines JL, Pericak-Vance MA, Schellenberg G, Kauwe JSK, Lunetta KL, Farrer LA, & Alzheimer's Disease Sequencing Project. (2019). Association of rare coding mutations with Alzheimer's disease and other dementias among adults of European ancestry. JAMA Netw Open, 2, e191350. *doi:* 10.1001/jamanetworkopen.2019.1350 [PubMed: 30924900]
34. Miller JB, Ward E, Staley LA, Stevens J, Teerlink CC, Tavana JP, Cloward M, Page M, Dayton L, Alzheimer's Disease Genetics Consortium, Cannon-Albright LA, & Kauwe JSK (2020). Identification and genomic analysis of pedigrees with exceptional longevity identifies candidate rare variants. Neurobio Dis, 143, 104972. *doi:* 10.1016/j.nbd.2020.104972
35. Thompson BA, Snow AK, Koptiuch C, Kohlmann WK, Mooney R, Johnson S, Huff CD, Yu Y, Teerlink CC, Feng BJ, Neklason DW, Cannon-Albright LA, & Tavtigian SV (2020). A novel ribosomal protein S20 variant in a family with unexplained colorectal cancer and polyposis. Clin Genet, 97, 6. *doi:* 10.1111/cge.13757
36. Cannon-Albright LA, Farnham JM, Stevens J, Teerlink CC, Palmer CA, Rowe K, Cessna MH, & Blumenthal DT (2020). Genome-wide analysis of high-risk primary brain cancer pedigrees identifies PDXDC1 as a candidate brain cancer predisposition gene. Neuro-Oncology, noaa161. *doi:* 10.1093/neuonc/noaa161
37. Cannon-Albright LA, Teerlink CC, Stevens J, Snow AK, Thompson BA, Bell R, Nguyen KN, Sargent NR, Kohlmann W, Neklason DW, & Tavtigian SV (2020). FANCM c5791CT stopgain mutation (rs144567652) is a familial colorectal cancer risk factor. Molec Genet & Genomic Med, e1532. *doi:* 10.1002/mgg3.1532 [PubMed: 33118316]
38. Teerlink CC, Jurynec MJ, Hernandez R, Stevens J, Hughes DC, Brunker CP, Rowe K, Grunwald DJ, Facelli JC, Cannon-Albright LA (2020). A role for the MEGF6 gene in predisposition to osteoporosis. Annals of Hum Genet, 12408. *doi:* 10.1111/ahg.12408
39. The GTEx Consortium. (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet, 45, 580–585. *doi:* 10.1038/ng.2653 [PubMed: 23715323]
40. ENCODE Project Consortium. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature, 489, 57–74. *doi:* 10.1038/nature11247 [PubMed: 22955616]
41. Glisovic T, Bachorik JL, Yong J, & Dreyfuss G (2008). RNA-binding proteins and posttranscriptional gene regulation. FEBS Lett, 582, 1977–1986. *doi:* 10.1016/j.febslet.2008.03.004 [PubMed: 18342629]
42. Wu YQ, Ju CL, Wang BJ, & Wang RG (2019). PABPC1L depletion inhibits proliferation and migration via blockage of AKT pathway in human colorectal cancer cells. Oncol Lett, 17, 3439–3445. *doi:* 10.3892/ol.2019.9999 [PubMed: 30867782]
43. Yu C, Yu J, Yao X, Wu WKK, Lu Y, Tang S, Li X, Bao L, Li X, Hou Y, Wu R, Jian M, Chen M, Zhang F, Xu L, Fan F, He J, Liang Q, Wang H, Hu X, He M, Zhang X, Zheng H, Li O, Wu H, Chen Y, Yang X, Zhu S, Xu X, Yang H, Wang J, Zhang X, Sung JJY, Li L, & Jun Wang. (2014).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Discovery of biclonal origin and a novel oncogene SLC12A5 in colon cancer by single-cell sequencing. Cell Res, 24, 701–712. *doi:* 10.1038/cr.2014.43 [PubMed: 24699064]

44. Yuan L, Xiao Y, Zhou Q, Yuan D, Wu B, Chen G, & Zhou J (2014). Proteomic analysis reveals that MAEL, a component of nuage, interacts with stress granule proteins in cancer cells. Oncol Rep, 31, 342–350. *doi:* 10.3892/or.2013.2836 [PubMed: 24189637]
45. Poulogiannis G, Ichimura K, Hamoudi RA, Luo F, Leung SY, Yuen ST, Harrison DJ, Wyllie AH, & Arends MJ (2010). Prognostic relevance of DNA copy number changes in colorectal cancer. J Pathol, 220, 338–347. doi: 10.1002/path.2640 [PubMed: 19911421]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

HIGHLIGHTS

- A pedigree-based approach identified a colorectal cancer risk variant in CELF4
- The variant was replicated using a population-based approach
- RNA structure analysis predicted the variant is destabilizing to the RNA structure
- Loss of CELF4 is a previously suggested prognostic indicator of colorectal cancer

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

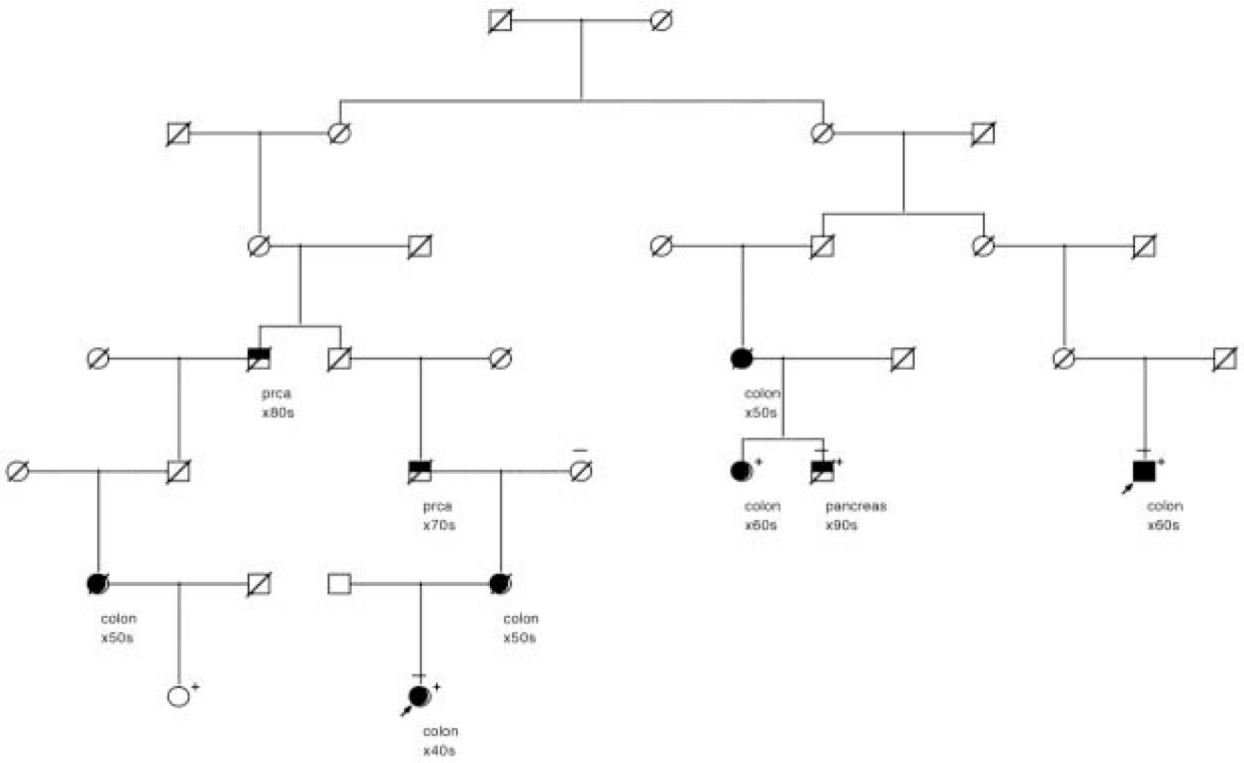

Figure 1.

Pedigree linked to chromosome 18q12.2. Subjects with colorectal cancer have dark fill, subjects with other cancers have half shading. Subjects with genotype data who were initially analyzed for linkage appear with a "−". Arrows indicate subjects that were selected for whole genome sequencing. Taqman-confirmed carriers of the *CELF4* variant are denoted with a '+'. 'Prca' denotes prostate cancer.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

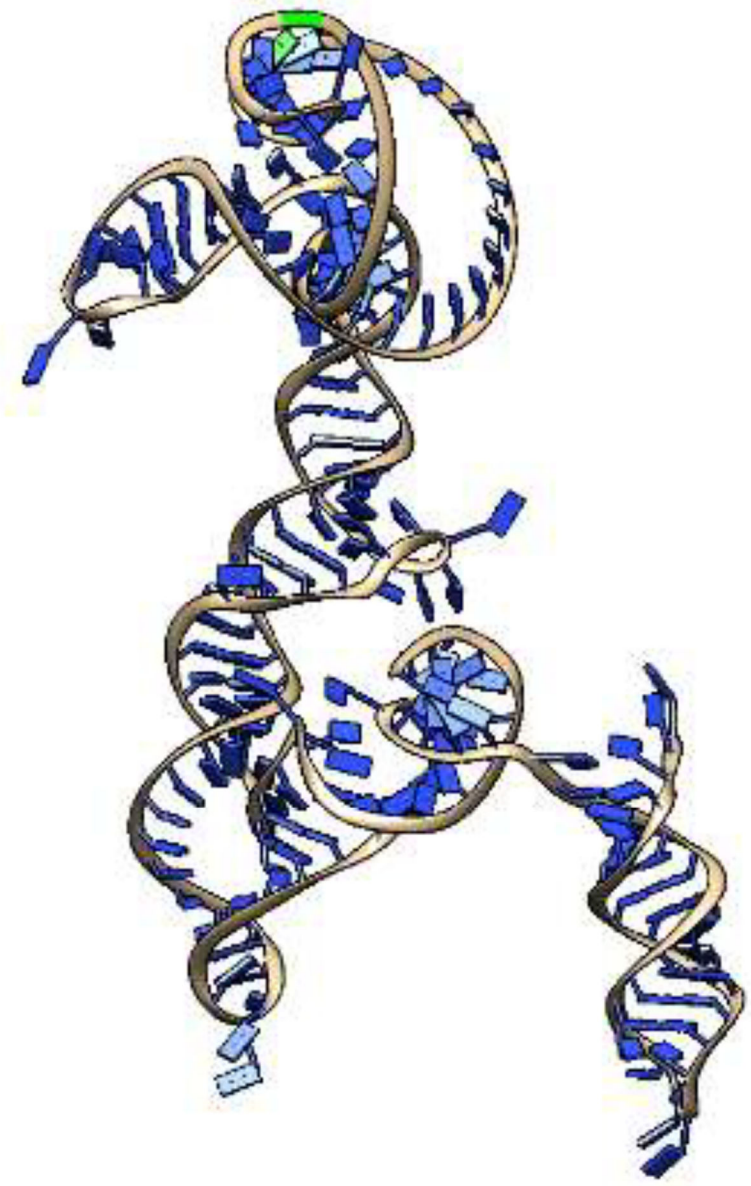

Figure 2.
Wild type RNA structure. The position affected in the variant is highlighted in green. Both the backbone and nucleic acids are shown.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

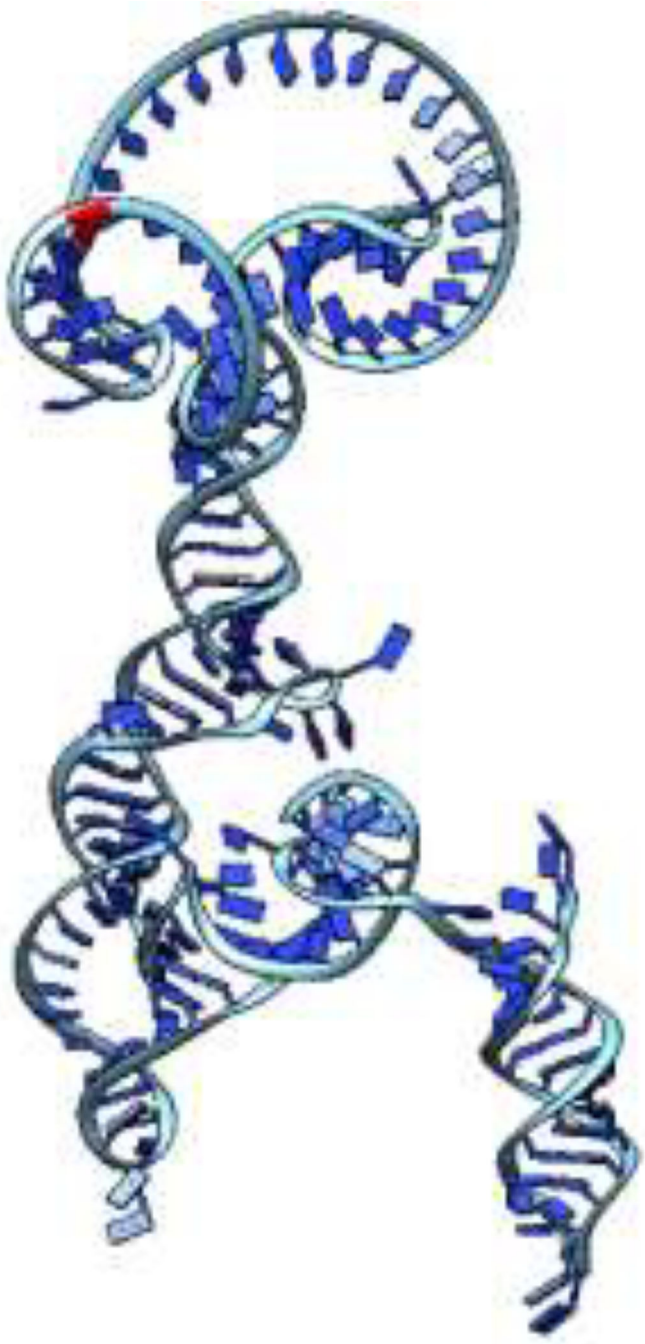

Figure 3.
Variant RNA structure. The affected position is highlighted in red. Both the backbone and nucleic acids are shown.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

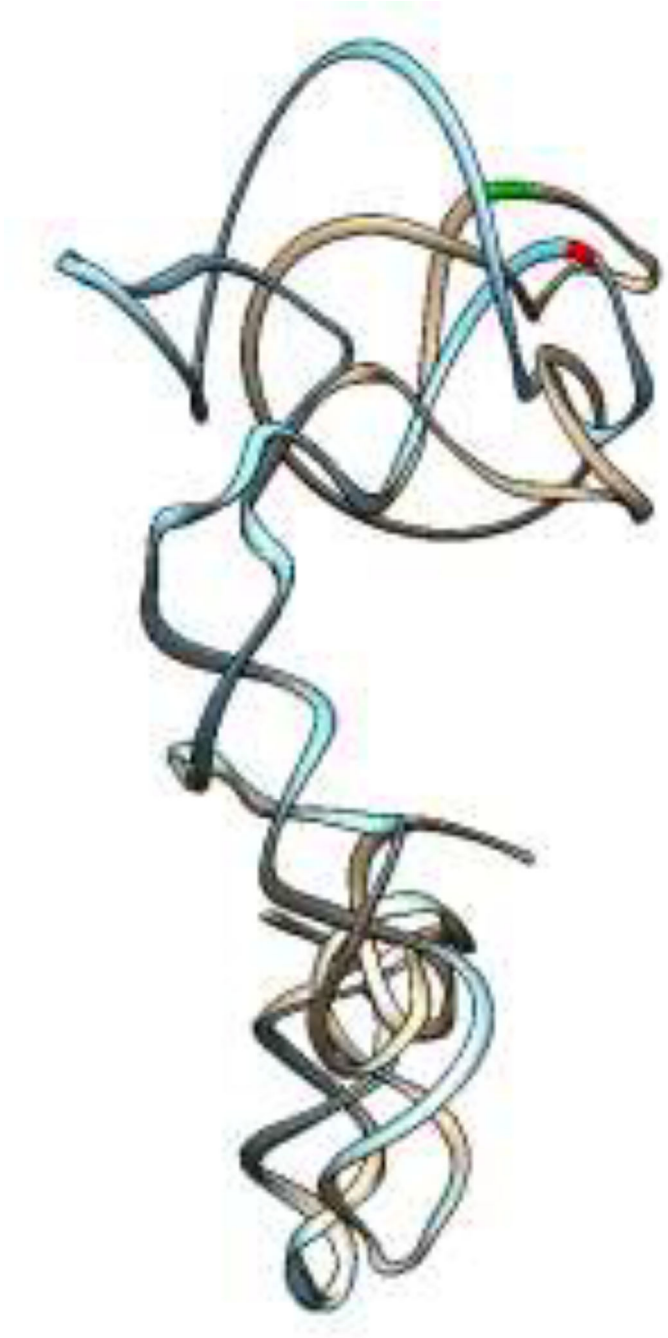

Figure 4.
Wild type and variant RNA structures superimposed. The conformational change is visible at the top of the image. Only backbones shown.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Figure 5.
Wild type and variant RNA structures superimposed. The conformational change is visible at the top of the image. Backbones and nucleic acids are shown.