nature genetics

The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling

Sari Tuupanen[1], Mikko Turunen[2,3], Rainer Lehtonen[1], Outi Hallikas[2,3], Sakari Vanharanta[1,12], Teemu Kivioja[2–4], Mikael Björklund[2,3], Gonghong Wei[2,3], Jian Yan[2,3], Iina Niittymäki[1], Jukka-Pekka Mecklin[5], Heikki Järvinen[6], Ari Ristimäki[7–9], Mariachiara Di-Bernardo[10], Phil East[11], Luis Carvajal-Carmona[11], Richard S Houlston[10], Ian Tomlinson[11], Kimmo Palin[4,12], Esko Ukkonen[4], Auli Karhu[1], Jussi Taipale[2,3] & Lauri A Aaltonen[1]

Homozygosity for the G allele of rs6983267 at 8q24 increases colorectal cancer (CRC) risk ~1.5 fold. We report here that the risk allele G shows copy number increase during CRC development. Our computer algorithm, Enhancer Element Locator (EEL), identified an enhancer element that contains rs6983267. The element drove expression of a reporter gene in a pattern that is consistent with regulation by the key CRC pathway Wnt. rs6983267 affects a binding site for the Wnt-regulated transcription factor TCF4, with the risk allele G showing stronger binding *in vitro* and *in vivo*. Genome-wide ChIP assay revealed the element as the strongest TCF4 binding site within 1 Mb of *MYC*. An unambiguous correlation between rs6983267 genotype and *MYC* expression was not detected, and additional work is required to scrutinize all possible targets of the enhancer. Our work provides evidence that the common CRC predisposition associated with 8q24 arises from enhanced responsiveness to Wnt signaling.

Genome-wide association studies (GWAS) are a powerful way to identify disease susceptibility variants[1,2]. However, the molecular bases of the observed associations often remain unclear because the markers used merely serve as indicators that a causative variant is present nearby. Pinpointing these changes is a major challenge in unravelling the genetic basis of complex disease predisposition.

The SNP rs6983267 at chromosome 8q24.21 has been implicated in predisposition to colorectal (CRC) and prostate cancer[3–7]. The CRC association has been confirmed in several studies[8–12]. The risk extends also to benign colonic adenomas, indicating that the causative variant already has a role at the precancerous stage[3,13]. The risk allele G is common, with an allele frequency of 50% in individuals of European descent and almost 100% in populations with African origin[5]. Thus, although the increase in CRC risk is modest (approximately 1.5-fold in GG homozygotes[3,10]), the contribution of rs6983267 to CRC incidence globally is extremely important.

The mechanism by which this SNP increases risk of CRC has not been elucidated, and resequencing of the region has not revealed obvious pathogenic changes[14]. The nearest protein-coding gene, *MYC*, resides 335 kb downstream from rs6983267; hence, one hypothesis has been that the SNP might be in linkage disequilibrium (LD) with unknown elements controlling the expression of this key oncogene. Possible existence of a regulatory region at rs6983267 has been proposed according to the UCSC genome browser, which indicates that this region is highly conserved during evolution, and the VISTA enhancer database, which predicts a putative enhancer in the region (see URLs section in Online Methods)[14]. The vicinity of *MYC* is of interest as this gene is a major CRC oncogene and one of the most well-established targets of Wnt signaling pathway[15]. The Wnt signaling pathway is aberrantly activated in over 90% of CRCs, most commonly through mutations in the *APC* (adenomatous polyposis coli) gene[16]. In the absence of Wnt ligand, APC participates in the degradation of the central cytoplasmic signal transducer of the Wnt pathway, β-catenin. Binding of Wnt proteins to their receptors on the cell surface results in stabilization of β-catenin, which enters the nucleus and functions as cofactor for transcription factors of the TCF

[1]Department of Medical Genetics and [2]Institute of Biomedicine, Genome-Scale Biology Research Program, Biomedicum Helsinki, University of Helsinki, Helsinki, Finland. [3]Department of Molecular Medicine, National Public Health Institute. [4]Department of Computer Science, University of Helsinki, Helsinki, Finland. [5]Department of Surgery, Jyväskylä Central Hospital, Jyväskylä, Finland. [6]The Second Department of Surgery, Helsinki University Hospital, Helsinki, Finland. [7]Genome-Scale Biology Research Program, University of Helsinki, Helsinki, Finland. [8]Department of Pathology, HUSLAB and Haartman Institute, Helsinki University Central Hospital, Helsinki, Finland. [9]Department of Pathology, University of Oulu and Oulu University Hospital, Oulu, Finland. [10]Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey, UK. [11]Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Headington, Oxford, UK. [12]Present addresses: Cancer Biology and Genetics Program, Memorial Sloan-Kettering Cancer Center, New York, New York, USA (S.V.) and Wellcome Trust Sanger Institute, Hinxton, Cambridge, UK (K.P.). Correspondence should be addressed to L.A.A. (lauri.aaltonen@helsinki.fi) or J.T. (jussi.taipale@helsinki.fi).

Received 5 March; accepted 20 May; published online 28 June 2009; doi:10.1038/ng.406

(T-cell factor)/LEF (lymphoid enhancer factor) family. Tumor-promoting mutations inactivating *APC* hamper degradation of β-catenin and lead to constitutive ligand-independent activation of TCF/LEF-mediated transcription. Germline variants affecting responsiveness of key enhancer elements to TCF/LEF could play an important role in CRC susceptibility.

We have previously shown that somatically emerging allelic imbalance at rs6983267 in tumor DNA favors the G risk allele[10]. We did not examine whether the imbalance is due to loss of the nonrisk T allele or gain of the risk G allele. The former would be compatible with the classical tumor suppressor mechanism typically seen in high-penetrance hereditary CRC syndromes, whereas the latter would suggest an oncogenic change. In the present study, we aimed to shed light to the mechanisms that underlie CRC susceptibility associated with rs6983267.

RESULTS

Copy number analysis in CRC tumors

In our previous work[10], we had identified 67 CRCs in which the G allele was selected for, as compared with 34 with T allele selection, showing that the G allele is favored during genesis of CRC ($P = 0.0007$). To characterize the nature of G allele overrepresentation, we analyzed paired normal and tumor DNAs from 12 heterozygous CRC cases (**Supplementary Note**) showing an excess of G allele as compared with T allele in cancer DNA using the Illumina HumanCNV370-Duo SNP array. Eleven out of 12 tumors showed copy number increase of a large segment of the G allele–bearing chromosome region (typically 3 to 5 copies) (**Fig. 1**). In nine tumors, the amplicon included the whole long arm of chromosome 8 (**Fig. 1a**). Two samples showed gains of shorter chromosome 8 segments (**Fig. 1b**). In one case, the experiment was not conclusive, although additional SNP sequencing analysis gave evidence for an imbalance spanning at least from rs10505476 (chr. 8:128,477,298 bp) to rs10505474 (chr. 8:128,486,686 bp). No cases with T-allele loss were identified, establishing copy number increase as the cause for preferred overrepresentation of the G allele in CRCs ($P = 0.001$). This result strongly suggests that the function of the causative germline alteration is likely to be oncogenic.

Imputing and association analyses

To further investigate the 8q24 locus association and estimate the fraction of common variation at this locus, we imputed genotype data from the GWAS of Tomlinson *et al.*[3]. In total, 172 HapMap SNPs were successfully imputed in the interval between 128.419 Mb and 128.562 Mb at 8q24 using the Illumina-generated SNP genotype data from CORGI (29 SNPs from the Illumina Hap550 bead array) and NSCCG1 (83 SNPs from an Illumina Infinium custom iSelect array) (**Supplementary Note** and **Supplementary Fig. 1**). rs6983267 showed the peak of the association signal in the combined analysis ($P = 1.16 \times 10^{-10}$). Second to that was rs10505477 ($P = 6.41 \times 10^{-10}$) and the third-ranked SNP was rs7014346 ($P = 2.47 \times 10^{-7}$), identified as the top SNP at the locus by Tenesa *et al.*[17]. rs10505477 and rs698327 have similar allele frequency and are strongly correlated ($r^2 = 0.92$). Owing to few recombination events within the 5.86-kb region of LD containing rs10505477 and rs693267, there are effectively two major 8q24.21 haplotypes in the disease, compatible with a single disease-causing variant tagged by rs693267. This conclusion is also compatible with data reported by Yeager *et al.*[14], who resequenced the region in 79 prostate cancer cases and controls of European origin and were unable to find new potentially pathogenic variants.

Prediction of enhancer elements around *MYC*

To examine the nature of the putative oncogenic change at the rs6983267 region, we used the computational method Enhancer Element Locator (EEL)[18] to detect possible regulatory motifs. EEL predicts functional enhancer elements by aligning genomic regions from two different species according to the order of transcription factor binding sites that they encode. It scores the elements on the basis of transcription factor affinities and clustering and conservation of the sites (for details, see refs. 18,19). We used 1 Mb of mouse and

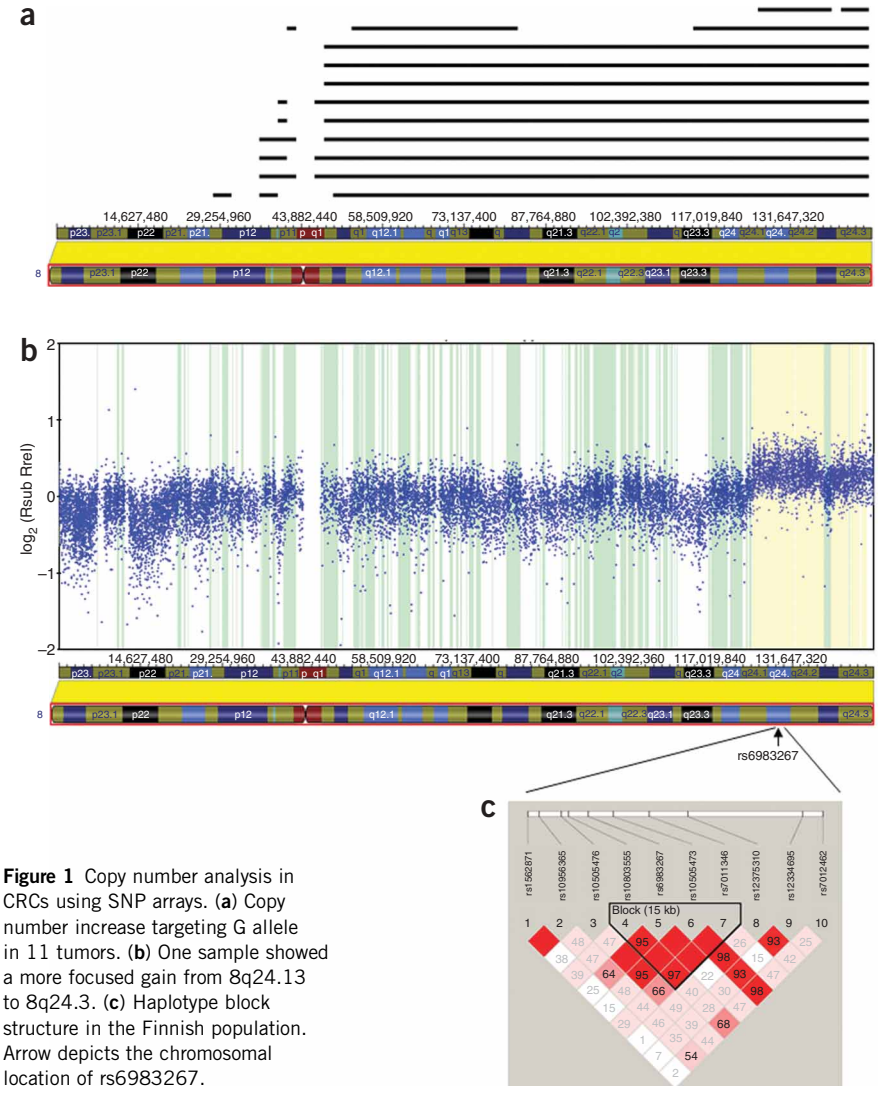

Figure 1 Copy number analysis in CRCs using SNP arrays. (**a**) Copy number increase targeting G allele in 11 tumors. (**b**) One sample showed a more focused gain from 8q24.13 to 8q24.3. (**c**) Haplotype block structure in the Finnish population. Arrow depicts the chromosomal location of rs6983267.

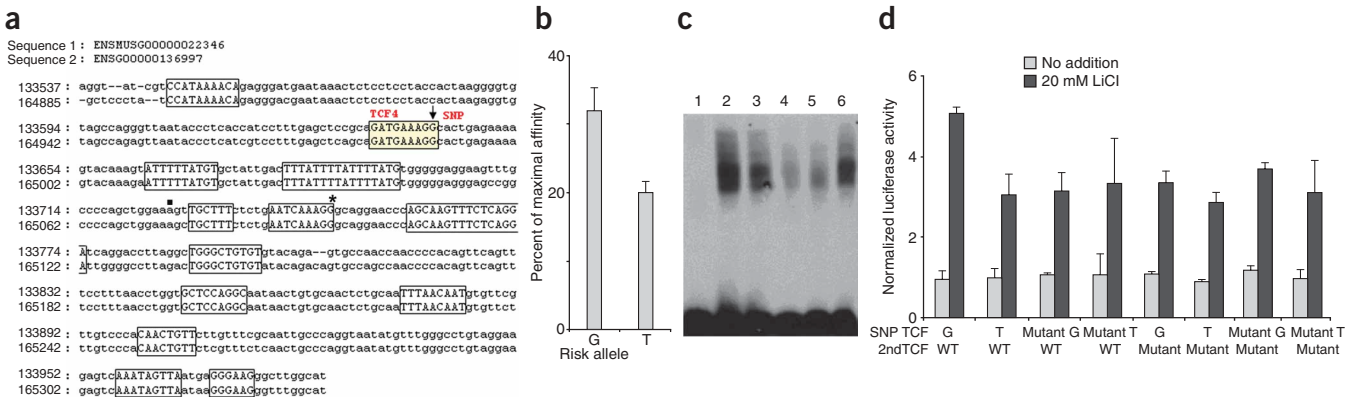

Figure 2 rs6983267 maps to a TCF4 site at a conserved regulatory element. (**a**) EEL alignment predicts multiple conserved transcription factor binding sites in the ~0.5-kb region, which is located 335 kb upstream of *MYC* and contains rs6983267 (upper panel). Mouse and human sequences of the predicted enhancer element are on top and bottom lines, respectively (lower panel). Yellow box indicates a conserved TCF4 site whose affinity is altered by rs6983267 (arrow). Asterisk indicates a second TCF4 site. The black square depicts SNP rs34835043, which is not predicted to affect the functionality of the enhancer element and has a very low minor-allele frequency. (**b**) Relative TCF4 affinities of sequences containing risk (G) and nonrisk (T) alleles of rs6983267, as compared with consensus sequence. The relative sequence-specific binding affinities were measured using the luciferase-based *in vitro* binding assay[18]. Error bars, s.e.m. ($n = 5$). (**c**) The G allele shows higher affinity for TCF4 than the T allele in EMSA. Lane 1 represents negative control. The binding of TCF4 consensus sequence (lane 2) is competed by an unrelated sequence (lane 3), by the TCF4 consensus sequence (lane 4) and by the sequence containing the rs6983267 G allele (lane 5) and T allele (lane 6). (**d**) MYC-335 shows Wnt-responsive enhancer activity in luciferase assay in HEK293T cells. After activation of the Wnt pathway by LiCl, the MYC-335 reporter containing the G allele of rs6983267 shows stronger induction than a similar reporter containing the T allele. Error bars, s.d. ($n = 4$). SNP TCF, TCF4 binding site containing rs6983267; 2nd TCF, TCF4 site marked with asterisk in **a**.

human genomic sequences around *MYC* to predict conserved enhancer elements (see Online Methods). The strongest EEL-predicted regulatory element within this large region contained the CRC predisposition SNP rs6983267 (**Fig. 2a**, **Supplementary Fig. 2** and **Supplementary Table 1**). To determine whether one or more of the other, weaker predicted regulatory elements also reside in LD with rs6983267, we determined the haplotype block structure in the rs6983267 region in Finns by Haploview 4.0 (ref. 20) using HumanHap300 (Illumina) SNP data from 265 Finnish control individuals. rs6983267 was found to reside in a 15,281-bp haploblock flanked by SNPs rs10808555 (chr. 8:128,478,693 bp) and rs7014346 (chr. 8:128,493,974 bp) (**Fig. 1c**). The strong predicted element that contains rs6983267 was the only element predicted by EEL within the haploblock (EEL score > 100; a cutoff that gives >17 elements per aligned gene; **Supplementary Table 1** and data not shown), suggesting that rs6983267 could be the causative change.

Affinity analyses of the predicted TCF4 site

Notably, EEL predicted that rs6983267 directly affected a binding site for TCF4 (also known as TCF7L2), a transcription factor activated in most CRCs[16] (**Fig. 2a** and **Supplementary Fig. 2**). The TCF4 binding sequence 5′-GATGAAAGG-3′ identified here matches closely to the TCF4 consensus sequence CATCAAAGG[18]. We previously created a binding affinity matrix for TCF4, and according to these data the change from T to G at the last position (5′-GATGAAAGT-3′ to 5′-GATGAAAGG-3′) considerably increases the affinity of TCF4 for this site (ref. 18 and data not shown). As the TCF4 site identified here is not the exact consensus sequence, we specifically tested the affinity of the exact sequences corresponding to the T and G allele, GATGAAAGT and GATGAAAGG, and found that they bound TCF4 *in vitro* with sequence-specific DNA-binding affinities that are 20% and 32% of the maximal affinity site, respectively (**Fig. 2b**). These affinities are well within the range reported for physiologically relevant transcription factor binding sites[21,22]. We also conducted electrophoretic mobility shift assays (EMSA), which confirmed that the G allele has higher affinity to TCF4 than the T allele (**Fig. 2c**). Another TCF4 site found in the predicted element, AATGAAAGG, has earlier been reported to have approximately 12% of maximal affinity[18]. No other conserved TCF4 sites with significant predicted activity lie within this enhancer.

The element containing rs6983267 is a Wnt-responsive enhancer

To test whether the predicted element (which we refer to hitherto as MYC-335) could function as a Wnt-pathway–dependent enhancer element in cultured cells, we cloned MYC-335 elements containing the T and G alleles of rs6983267 into a reporter vector containing a minimal promoter followed by a firefly luciferase reporter gene. The promoter is inactive in the absence of enhancer activity derived from the cloned fragments and can thus be used to test whether MYC-335 has enhancer activity in cultured cells. Transfection of these vectors into Wnt-responsive HEK293T cells in the absence of Wnt induction revealed no significant change in reporter activity between the G and T allele (**Fig. 2d**). However, activation of the Wnt pathway through GSK3β inhibition by lithium chloride revealed that MYC-335 has Wnt-responsive enhancer activity, and that the G allele shows 1.5-fold stronger Wnt responses than the T allele. Mutation of the second and third nucleotides of the TCF4 sites in the G and T allele reporters resulted in partial loss of Wnt responsiveness and complete loss of the effect of the SNP, indicating that the observed increase in the induction of the G allele was dependent on the TCF4 site containing rs6983267. We also mutated the other TCF4 site individually and together with the SNP-containing TCF4 site. All mutants were similar to the weaker T allele in their response to Wnt signaling, indicating that both TCF4 sites are required for the 1.5-fold increased induction in the G allele–containing reporter construct (**Fig. 2d**).

The rs6983267 site is occupied by TCF4 and β-catenin in CRC cells

To assess whether the MYC-335 enhancer element is active in its normal chromosomal context in CRC cells, we conducted a chromatin immunoprecipitation (ChIP) assay (**Fig. 3**). First, 12 CRC cell lines

were genotyped for rs6983267. VACO5, GP5D, HCT8, DLD1/HCT15, LoVo, LS174T and LS180 are heterozygous at rs6983267, SW480, HCA7, HCT116, HUTU80 and RKO are homozygous for G and CCL-231 is homozygous for T. ChIP assays were initially done in LoVo cells using antibodies to TCF4, β-catenin and control IgG (**Fig. 3a**,**b**). These experiments showed that the rs6983267-containing region indeed binds the DNA-binding protein TCF4 and the transcriptional activator β-catenin *in vivo*. Strong binding of TCF4 to the site was confirmed in all six other heterozygous cell lines tested, as well as in CCL-231 (TT) (**Fig. 3c**). Negative control regions did not show significant binding of TCF4 in any of the cell lines tested (**Fig. 3a**,**b** and data not shown).

We sequenced input and anti-TCF4 immunoprecipitated DNA, and then analyzed the signal intensity for the individual alleles (**Fig. 3f**,**g**)[23]. This analysis on LoVo cells revealed that the risk-allele G at the site enhances TCF4 binding approximately twofold (**Fig. 3f**), consistent with our earlier *in vitro* binding-affinity experiments (**Fig. 2b**)[18]. The preferential binding of the G allele to TCF4 was also observed by sequencing the TCF4 immunoprecipitated rs6983267-containing PCR product in all other cell lines heterozygous for rs6983267 (**Fig. 3g**).

To further validate the MYC-335 element, we systematically identified TCF4 binding sites across the entire human genome both experimentally and computationally using chromatin immunoprecipitation by sequencing (ChIP-Seq)[24] in LoVo cells and EEL alignment[18], respectively. Integration of these analyses identified MYC-335 as the fourth-highest ranked element containing a conserved TCF4 site in the entire human genome (**Table 1**). Nine out of ten highest-ranked elements identified were located close to known Wnt/TCF4 target genes (**Table 1**). Furthermore, ChIP-Seq in LoVo cells revealed that TCF4 binds MYC-335 stronger than any other sequence within 1 Mb of *MYC* (**Fig. 3d**). Detailed analysis of the sequenced fragments indicated that the peak of binding is located at rs6983267, with 14 sequenced fragments containing only the TCF4 site at rs6983267 (**Fig. 3e** and **Supplementary Fig. 3**). Ten sequenced fragments (**Supplementary Fig. 3**) contained only the weaker TCF4 site located 106 bp downstream (**Fig. 2a**, asterisk), suggesting that, consistent with the luciferase reporter analyses, this site is also occupied in CRC cells. Strong binding of TCF4 to the SNP region was similarly detected in another colon cancer cell line, GP5D, using ChIP-seq (**Supplementary Fig. 4**). Furthermore, in both LoVo and GP5D cells, the region near rs6983267 was also enriched in lysine 4 monomethylated histone H3 (data not shown), a known marker for enhancer elements[25].

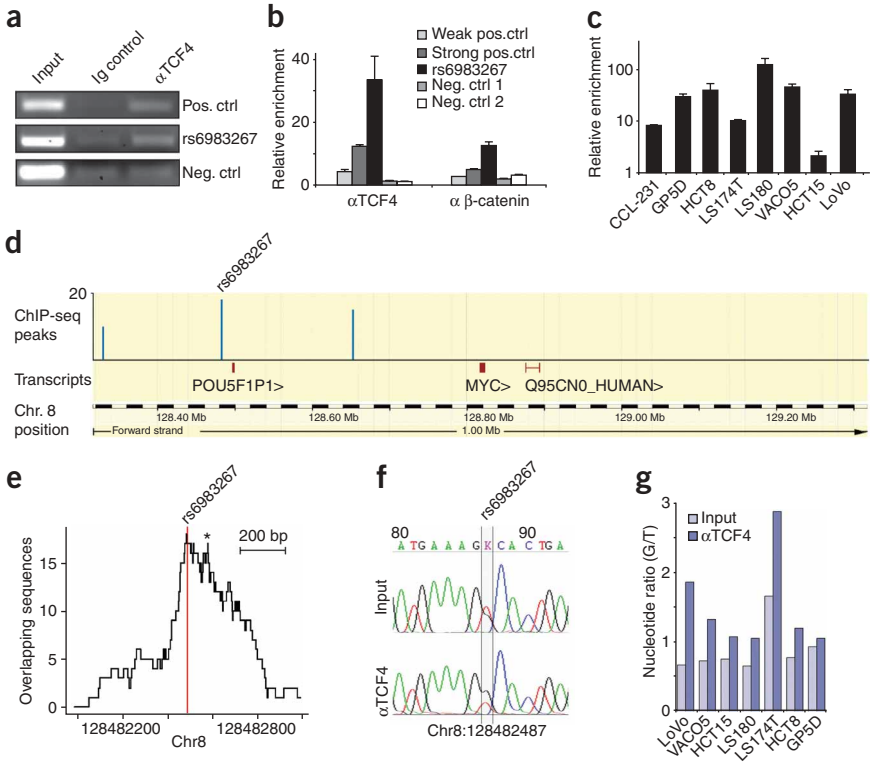

Figure 3 TCF4 and β-catenin occupancy at MYC-335 in CRC cells. (**a**) TCF4 binds to a DNA fragment containing rs6983267 in ChIP analysis in LoVo cells. PCR from input DNA (Input) and precipitates using anti-TCF4 (αTCF4) and control IgG (Ig control) are shown for a positive region (positive control, chr. 7:5890255L, ref. 23), rs6983267 and negative region (negative control 2; ref. 23). (**b**) ChIP and quantitative real-time PCR confirm that the rs6983267 site binds to both TCF4 and β-catenin in LoVo cells. Fold enrichments relative to IgG control are shown. (**c**) Confirmation of TCF4 binding to rs6983267 in eight different CRC cell lines by ChIP and quantitative real-time PCR. (**d**) ChIP-Seq analysis of 1 Mb flanking *MYC* showing peaks with height ≥10 and $P < 0.05$. (**e**) The 1-kb region with the highest number of overlapping sequences as determined by ChIP-Seq. rs6983267 resides at the peak. An asterisk depicts a shoulder peak at the other TCF4 site (**Fig. 2a**). (**f**) ChIP followed by PCR and Sanger sequencing for the rs6983267-containing region reveals that TCF4 favors binding to the risk allele of rs6983267 in LoVo cells. Sanger-sequencing of anti-TCF4–precipitated DNA (αTCF4) reveals that the risk allele (G) is enriched by the TCF4 precipitation (bottom panel) compared to input DNA (top panel). (**g**) Sanger sequencing analysis of G to T nucleotide ratios at rs6983267 site in seven heterozygous cell lines before (light bars) or after (dark bars) anti-TCF4 ChIP. TCF4 preferentially binds to the G allele in all cases.

MYC-335 functions as a tissue-specific enhancer *in vivo*

To test whether the MYC-335 element also functions as an enhancer element *in vivo*, we cloned the element in front of a *LacZ* reporter gene and injected the construct into mouse oocytes. In the resulting embryos, MYC-335 drove *LacZ* expression in the tail and/or limbs at embryonic day 12.5 (E12.5) in seven of eight embryos with *LacZ* staining, verifying the functionality of the element *in vivo*. The observed expression pattern is consistent with that of Wnt signaling target genes[26,27] (**Fig. 4**, left embryo). Consistently, we observed only ectopic *LacZ* staining in embryos injected with a construct where the two conserved TCF4 sites in MYC-335 (**Fig. 2a**) had been mutated (**Supplementary Fig. 5**). The MYC-335–driven *LacZ* is expressed in a subset of the tissues where *Myc* RNA is expressed (**Fig. 4**, right embryo).

Correlation between rs6983267 genotype and *MYC* expression

We examined a possible correlation between rs6983267 genotype and *MYC* expression in lymphoblasts, normal colon epithelial cells and CRCs. Only a marginally significant relationship between genotype and gene-level (16 probesets) expression of *MYC* was observed in lymphoblasts (empirical Bayes statistics, $P = 0.04$, not corrected for

Table 1 Genome-wide analysis of TCF4 binding sites

Peak no.	Conserved gene(s) whose transcription start site is within 500 kb of the indicated Tcf4 ChIP-Seq peak	Peak height	Chr.	Peak start	Peak end	Peak *P* value	EEL score
1	*NR2F2*	29	15	95038245	95038719	1.8×10^{-8}[a]	1927
2	*PAPSS1*,***LEF1***[38],*SGMS2*,*CYP2U1*	26	4	109313896	109314226	1.4×10^{-5}[a]	419
3	***AXIN2***[38],*RGS9*,*CCDC46*	24	17	60987683	60987875	1.7×10^{-5}[a]	484
4	***MYC***[38]	18[b]	8	128482240	128482827	5.8×10^{-3}[a]	323
5	*C10orf107*,*PLEKHK1*,***ARID5B***[38],*TMEM26*	16	10	63267081	63267277	5.2×10^{-3}[a]	632
6	***HOXC4***[39], *HOXC6*,*HOXC8–13*, *CALCOCO1*,*ATF7*,*NPFF*,*SMUG1*, *ITGA5*,*COPZ1*, *GPR84*,*ATP5G2*,*TARBP2*,*FAM112B*, *CBX5*,*ZNF385*,*MAP3K12*,*NFE2*,*HNRNPA1*	13	12	52659125	52659334	1.1×10^{-1}	762
7	***AXIN2***[38],*APOH*,*CCDC46*	13	17	61177211	61177541	1.8×10^{-3}[a]	352
8	*HOXA1–7*,*HOXA9*,***HOXA10***[40],***HOXA11***[38,40],*HOXA13*,*TAX1BP1*,*SKAP2*,*JAZF1*, *HIBADH*,*EVX1*	12	7	27351746	27351926	2.1×10^{-3}[a]	814
9	*XRN2*,***C20orf19***[38],*NKX2-4*,*C20orf74*,***NKX2-2***[41]	12	20	20946395	20946628	2.7×10^{-3}[a]	729
10	*ENPEP*,***PITX2***[42]	12	4	112067593	112067774	2.1×10^{-3}[a]	305

Genome-wide analysis by ChIP sequencing shows that the rs6983267-containing regulatory element (MYC-335) in the vicinity of the *MYC* gene is the fourth-highest-ranked conserved TCF4-bound regulatory element in LoVo CRC cells. Ten highest-scoring conserved Tcf4-site-containing regulatory elements (by ChIP-Seq peak height) identified by a combination of genome-wide ChIP-Seq and EEL analyses (human to mouse comparison). Genes located within 500 kb of the elements, height of the ChIP-Seq peak and its chromosomal position (NCBI36), *P* value (see Online Methods for details) and EEL score are shown. Known Wnt/TCF4 targets are indicated in bold.
[a]Statistically significant. [b]Contains rs6983267.

multiple testing), using publicly available exon array and genotype data from 57 CEU HapMap individuals (see Online Methods). By comparing *MYC* expression in normal colonic epithelium from 33 GG and 35 TT CRC samples, we observed no significant difference between the groups ($P = 0.83$, *t*-test). Similarly, examination of our previously published gene expression data on 34 CRCs did not show a significant dependence between rs6983267 genotype and *MYC* expression ($P = 0.48$). Indeed, no correlation between rs6983267 genotype and *MYC* expression levels has been observed in previous studies by others using lymphoblastoid cell lines[28] and CRCs[8].

DISCUSSION

Genome-wide association studies are identifying an increasing number of SNPs associated with disease susceptibility. The challenge is to unravel the causative mechanisms related to these variants. The first such study of CRC pinpointed six susceptibility loci: rs6983267 at 8q24 (ref. 3), rs4939827, rs12953717 and rs4464148 at 18q21 (*SMAD7*)[29], rs4779584 and rs10318 at 15q[30], rs3802842 at 11q23 (ref. 17), rs10795668 at 10p14 and rs16892766 at 8q23.3 (ref. 31). Recent meta-analysis has identified four new susceptibility loci at 14q22 (rs4444235), 16q22 (9929218), 19q13 (rs10411210) and 20p12 (rs961253)[28]. At 8q24, multiple different regions associate with different cancer subtypes, including colorectal[3], prostate[6,14] and bladder cancer[32]. The data in our study provide evidence that at least one mechanism involved is the potential to regulate gene expression. In addition to MYC-335, at least one other associated region at 8q24 contains regulatory elements. The region associated with bladder cancer[32] contains the *MYC* regulatory elements CM3 and CM5, which we identified previously[18]. MYC-335, CM3 and CM5 all have distinct tissue specificities, possibly explaining differences in associated 8q24 regions between cancer types.

Figure 4 Verification of the MYC-335 enhancer element in transgenic mouse embryos. Representative embryos are shown. Note that MYC-335 drives *LacZ* (left and middle embryo) in limbs (arrows) and tail (yellow arrowhead) in a pattern that is similar to endogenous *Myc* expression (detected by *in situ* hybridization; right embryo). The left embryo is E12.5, whereas the middle *LacZ* embryo is age-matched with the *in situ* embryo (E11.5).

The data presented in this study show that rs6983267 affects binding of the Wnt-regulated transcription factor TCF4 in a regulatory element that is functional in CRC cells. TCF4 is a member of a class of closely related Wnt-regulated transcription factors that includes TCF1, TCF3, TCF4 and LEF1. Although TCF4 is thought to be the key transducer of Wnt signals in CRC, it is possible that part of the increased responsiveness to Wnt signaling caused by the G allele of rs6983267 could also be transmitted through these other TCF family members.

The affinity difference between T and G alleles for TCF4 was found to be relatively small, and the magnitude of this effect is in proportion to the slightly increased—but on population level extremely important—CRC risk related to rs6983267 (ref. 10). Although the similarity of the magnitudes of effect is of note, it is clear that, in general, the size of the biological effect cannot be predicted by the epidemiological risk, or vice versa. It is also important to note that G is the ancestral allele, and in addition to conferring a risk to cancer may have yet unknown favorable effects.

Further work will unravel the target(s) of MYC-335. Despite the fact that we and others have been unable to robustly show association between rs6983267 genotypes and *MYC* expression, one cannot rule out the possibility that at least one of the targets may be *MYC*. Some considerations speak in favor of this option. First, *MYC* is the central oncogene whose expression in CRCs is upregulated by 8q24 amplification, and the simplest hypothesis explaining preferential amplification of the G risk allele is that this allele expresses more *MYC*. Therefore, its copy number increase would confer a stronger selective advantage to tumor cells than copy number increase of the protective allele T, which is also potentially tumorigenic, albeit slightly less so. Second, *MYC* is a known target of TCF4, and the MYC-335 element is the strongest TCF4-binding region within 1 Mb of *MYC*. Third, by generating enhancer *LacZ* transgenic mouse embryos, we showed that MYC-335 drives gene expression *in vivo* in a pattern that is compatible with that of Wnt-regulated genes such as *Myc*. Further evidence for *MYC* being a target of MYC-335 is provided in the accompanying paper by Freedman and colleagues[33], in which the rs6983267 region is

demonstrated to show long-range interaction with the *MYC* promoter, indeed suggesting a role for MYC-335 as a distant *MYC* enhancer.

Activation of the Wnt signaling pathway, including TCF4 as the major transcription factor, almost invariably occurs early during development of CRC[16]. This is usually achieved by inactivation of the *APC* tumor suppressor gene already in colonic adenoma formation, which leads to constitutively activated nuclear β-catenin/TCF4 transcription complex and activation of target gene expression such as *MYC*[15,34]. Myc is a crucial mediator of intestinal tumorigenesis in mice, as *Myc* deletion rescues the effect of *Apc* deficiency in the mouse intestine[35]. The excess of CRC and adenoma risk associated with the G allele of rs6983267, preferential somatic amplification of this allele during tumorigenesis, and the functional data presented here linking rs6983267 to TCF4 affinity is compatible with the extensive previous work identifying Wnt as a key player in genesis of colonic neoplasia. The possible existence of additional genetic variants at MYC-335 contributing to target gene expression cannot be excluded and should be further investigated. Although the importance of the Wnt pathway as a potential target for rational cancer therapies has been clear[36], these new results suggest that the same pathway may also have implications for the development of personalized cancer prevention strategies[37].

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

Accession codes. Ensembl: *Myc* input sequences, ENSG00000136997 and ENSMUSG00000022346. NCBI GEO: exon array used to correlate rs6983267 genotype to *MYC* expression, GSE9372; gene expression data from this study, GSE4045.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by the grants from Academy of Finland (Finnish Center of Excellence Program 2006-2011), the Finnish Cancer Society, the Sigrid Juselius Foundation, the European Commission (LSHG-CT-2004-512142) and by grants to S.T. (Ida Montin Foundation, Biomedicum Helsinki Foundation, Paulo Foundation, Mary and Georg C. Ehrnrooth Foundation and Maud Kuistila Foundation). The work of R.S.H. and I.T. is supported by Cancer Research UK. We thank A. Syvänen from the Uppsala University SNP Platform and O. Monni from the University of Helsinki for Illumina Genome-analyzer sequencing, G. Yochum for control primer sequences for TCF4, and S. Marttinen, K. Pylvänäinen, T. Lehtinen, S. Miettinen, M. Kuris, M. Aho and I. Svedberg for technical assistance. L. Peltonen (National Public Health Institute) and the Nordic Center of Excellence in Disease Genetics provided the Finnish control SNP data.

AUTHOR CONTRIBUTIONS

The study was designed and financial support was obtained by L.A.A. and J.T. The manuscript was drafted by L.A.A., J.T. and S.T. Wet-lab experiments were performed by S.T., M.T., O.H., M.B., G.W., J.Y. and I.N. M.D.-B., I.T. and R.S.H. provided the imputed SNP data. J.-P.M. and H.J. provided the Finnish CRC specimens, A.R. contributed to histopathological evaluation of materials. R.L., S.V., T.K., P.E., L.C.-C., K.P. and A.K. performed the computational and statistical analyses. K.P., E.U. and S.T. performed the EEL analyses.

Published online at http://www.nature.com/naturegenetics/.
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/.

1. Couzin, J. & Kaiser, J. Genome-wide association. Closing the net on common disease genes. *Science* **316**, 820–822 (2007).
2. Easton, D.F. & Eeles, R.A. Genome-wide association studies in cancer. *Hum. Mol. Genet.* **17**, R109–R115 (2008).
3. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* **39**, 984–988 (2007).
4. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
5. Haiman, C.A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nat. Genet.* **39**, 954–956 (2007).
6. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
7. Zheng, S.L. *et al.* Cumulative association of five genetic variants with prostate cancer. *N. Engl. J. Med.* **358**, 910–919 (2008).
8. Gruber, S.B. *et al.* Genetic variation in 8q24 associated with risk of colorectal cancer. *Cancer Biol. Ther.* **6**, 1143–1147 (2007).
9. Poynter, J.N. *et al.* Variants on 9p24 and 8q24 are associated with risk of colorectal cancer: results from the Colon Cancer Family Registry. *Cancer Res.* **67**, 11128–11132 (2007).
10. Tuupanen, S. *et al.* Allelic imbalance at rs6983267 suggests selection of the risk allele in somatic colorectal tumor evolution. *Cancer Res.* **68**, 14–17 (2008).
11. Li, L. *et al.* A common 8q24 variant and the risk of colon cancer: a population-based case-control study. *Cancer Epidemiol. Biomarkers Prev.* **17**, 339–342 (2008).
12. Schafmayer, C. *et al.* Investigation of the colorectal cancer susceptibility region on chromosome 8q24.21 in a large German case-control sample. *Int. J. Cancer* **124**, 75–80 (2009).
13. Berndt, S.I. *et al.* Pooled analysis of genetic variation at chromosome 8q24 and colorectal neoplasia risk. *Hum. Mol. Genet.* **17**, 2665–2672 (2008).
14. Yeager, M. *et al.* Comprehensive resequence analysis of a 136 kb region of human chromosome 8q24 associated with prostate and colon cancers. *Hum. Genet.* **124**, 161–170 (2008).
15. He, T.C. *et al.* Identification of c-MYC as a target of the APC pathway. *Science* **281**, 1509–1512 (1998).
16. Bienz, M. & Clevers, H. Linking colorectal cancer to Wnt signaling. *Cell* **103**, 311–320 (2000).
17. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.* **40**, 631–637 (2008).
18. Hallikas, O. *et al.* Genome-wide prediction of mammalian enhancers based on analysis of transcription-factor binding affinity. *Cell* **124**, 47–59 (2006).
19. Palin, K., Taipale, J. & Ukkonen, E. Locating potential enhancer elements by comparative genomics using the EEL software. *Nat. Protocols* **1**, 368–374 (2006).
20. Barrett, J.C., Fry, B., Maller, J. & Daly, M.J. Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics* **21**, 263–265 (2005).
21. Lickert, H. & Kemler, R. Functional analysis of *cis*-regulatory elements controlling initiation and maintenance of early *Cdx1* gene expression in the mouse. *Dev. Dyn.* **225**, 216–220 (2002).
22. Jiang, J. & Levine, M. Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen. *Cell* **72**, 741–752 (1993).
23. Yochum, G.S. *et al.* Serial analysis of chromatin occupancy identifies β-catenin target genes in colorectal carcinoma cells. *Proc. Natl. Acad. Sci. USA* **104**, 3324–3329 (2007).
24. Robertson, G. *et al.* Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat. Methods* **4**, 651–657 (2007).
25. Heintzman, N.D. *et al.* Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nat. Genet.* **39**, 311–318 (2007).
26. Barrow, J.R. *et al.* Ectodermal Wnt3/β-catenin signalling is required for the establishment and maintenance of the apical ectodermal ridge. *Genes Dev.* **17**, 394–409 (2003).
27. Huelsken, J. *et al.* Requirement for β-catenin in anterior-posterior axis formation in mice. *J. Cell Biol.* **148**, 567–578 (2000).
28. Houlston, R.S. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nat. Genet.* **40**, 1426–1435 (2008).
29. Broderick, P. *et al.* A genome-wide association study shows that common alleles of *SMAD7* influence colorectal cancer risk. *Nat. Genet.* **39**, 1315–1317 (2007).
30. Jaeger, E. *et al.* Common genetic variants at the *CRAC1* (*HMPS*) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nat. Genet.* **40**, 26–28 (2008).
31. Tomlinson, I.P. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nat. Genet.* **40**, 623–630 (2008).
32. Kiemeney, L.A. *et al.* Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nat. Genet.* **40**, 1307–1312 (2008).
33. Pomerantz, M.M. *et al.* The 8q24 cancer risk variant rs6983267 demonstrates long-range interaction with MYC in colorectal cancer. *Nat. Genet.* advance online publication, doi:10.1038/ng.403 (28 June 2009)..
34. Korinek, V. *et al.* Constitutive transcriptional activation by a β-catenin-Tcf complex in APC−/− colon carcinoma. *Science* **275**, 1784–1787 (1997).
35. Sansom, O.J. *et al.* Myc deletion rescues Apc deficiency in the small intestine. *Nature* **446**, 676–679 (2007).
36. Herbst, A. & Kolligs, F.T. Wnt signaling as a therapeutic target for cancer. *Methods Mol. Biol.* **361**, 63–91 (2007).
37. Soucek, L. *et al.* Modelling Myc inhibition as a cancer therapy. *Nature* **455**, 679–683 (2008).
38. Van der Flier, L.G. *et al.* The intestinal Wnt/TCF signature. *Gastroenterology* **132**, 628–635 (2007).
39. Otero, J.J., Fu, W., Kan, L., Cuadra, A.E. & Kessler, J.A. β-catenin signaling is required for neural differentiation of embryonic stem cells. *Development* **131**, 3545–3557 (2004).
40. Miller, C. & Sassoon, D.A. Wnt-7a maintains appropriate uterine patterning during the delopment of the mouse female reproductive tract. *Development* **125**, 3201–3211 (1998).
41. Lei, Q. *et al.* Wnt signaling inhibitors regulate the transcriptional response to morphogenetic Shh-Gli signaling in the neural tube. *Dev. Cell* **11**, 325–337 (2006).
42. Kioussi, C. *et al.* Identification of a Wnt/Dvl/β-catenin→Pitx2 pathway mediating cell-type-specific proliferation during development. *Cell* **111**, 673–685 (2002).

ONLINE METHODS

SNP array experiments and data analysis. Twelve tumor and matched normal DNAs from the Finnish CRC sample series[43,44] were hybridized to 24 Illumina HumanCNV370-Duo SNP array chips (Illumina) according to the manufacturer's instructions to examine whether the allelic imbalance in the tumor DNA observed in our previous study[10] was due to allelic loss or gain. CRC samples selected to the SNP arrays represented 60% or more carcinoma cells according to pathologist's evaluation. We analyzed copy number differences between tumors and respective normal samples with the BeadStudio 3.2 CNV Estimator tool using 0.2-Mb window size and 0.01 *P*-value cutoff.

Imputation and association analyses. Prediction of the untyped SNPs in the case-control datasets of NSCCG1 and CORGI (**Supplementary Note**) was carried out using MACH1.0 on HapMap (HapMap Data Rel 21a/phase II Jan07 on NCBI B35 assembly, dbSNPb125) phase II data. In total, 172 HapMap SNPs were successfully imputed in the interval between 128.419 Mb and 128.562 Mb at 8q24 using available SNP genotype data from NSCCG1 (83 SNPs) and CORGI (29 SNPs). We verified the integrity of imputed data where possible by crosschecking the concordance of imputed genotypes with that of available Illumina SNP genotype data. Concordance between imputed and directly measured genotypes was high. Specifically, for rs6983267 and rs10505477, concordance was greater than 99.6% (99.9% and 99.7%, respectively), thereby ensuring a high confidence in observations. Statistical analyses were undertaken in Stata Version 8 or R software. Linkage disequilibrium statistics were calculated using Haploview software (v4.0).

***In silico* methods.** To predict conserved enhancer elements, we aligned human and mouse genomic sequences 500 kb up- and downstream from *MYC* with the EEL algorithm essentially as described[18]. The *Myc* input sequences (chr. 8:128317498–129322853 and chr.15:61316896–62321907) were extracted from Ensembl database and masked for repetitive sequences using RepeatMasker program. The EEL analysis was run with default parameter values[19]. We used 107 high-quality transcription factor binding profiles were used[18].

For comparison of EEL results with TCF4 ChIP-Seq, we conducted a genome-scale EEL alignment essentially as described[18], with the following minor modifications: 160 transcription factor binding matrices (from Jaspar2; refs. 18,45) were used in alignment of 17,330 human genes with their mouse orthologs (from ensembl), including 500 kb of flanking sequences in both directions. Sequences were masked for tandem repeats (trf) and for known exons (ensembl 48). The aligned regions covered 80.6% of the human genome. The EEL alignment was executed with parameters lambda = 1.0, xi = 2684.59, mu = 0.144011 and nu = 2.98355, which were optimized with simulated annealing as previously described (see URLs section below). The binding site motif matrices had pseudocount of total 0.001 added according to the uniform, independent and identically distributed (i.i.d.) background. The sites with motif score ($\log_2$ odds between motif and uniform i.i.d. background) greater than 9 for at least one human haplotype were used in the alignment. The resulting data describing the scores and genomic locations of the predicted enhancer elements and the conserved transcription factor binding sites that they contain were loaded into a relational database. To generate **Table 1**, we ranked all TCF4 ChIP-Seq peaks that overlapped with predicted enhancer elements (EEL score > 300; this cut-off results in identification of approximately one regulatory element per aligned gene) and contained at least one conserved TCF4 site according to the height of the ChIP-Seq peak.

Cell culture. We obtained human CRC cell lines LoVo, HCT15, CCL-231, LS180, LS174T, VACO5, HCT8 and GPD5 from the American Type Culture Collection (ATCC) or the European Collection of Cell Cultures (ECACC). Cell culture was carried out using standard protocols.

Transcription factor binding assay. We conducted a luminescent-based transcription factor binding assay essentially as described[46], except that we took the 5-bp sequences flanking the TCF4 sites from the flanking sequences of the TCF4 binding site on MYC-335. This method is based on competition between binding sites and measures relative specific DNA-binding affinity of a transcription factor to a given sequence compared to the highest affinity sequence (relative specific affinity of 1) and to scrambled oligonucleotide control (relative specific affinity of 0). The double-stranded, blunt-ended oligonucleotide sequences used are described in **Supplementary Table 2**. The method has been extensively validated against EMSA previously using GLI2 and Tcf4 (ref. 18 and data not shown). The TCF4 binding affinity of MYC-335 TCF4 site was normalized as previously described[46].

Electrophoretic mobility shift assay (EMSA). EMSA was conducted using the LightShift Kit (Pierce). Double-stranded biotinylated TCF-consensus oligonucleotides were incubated for 1 h with control lysate or TCF4-*Renilla* luciferase fusion protein containing cell lysate[45] in the presence of no competitor or 50-fold molar excess of scrambled, TCF-consensus, G allele or T allele double-stranded competitor oligonucleotides (see **Supplementary Table 2** for oligonucleotide sequences). Binding buffer[45] was supplemented with 0.2% Triton X-100, 1% milk powder and 25 ng/µl poly(dI-dC). The resulting complexes were resolved in a 5% nondenaturing PAGE gel, transferred onto membrane and detected using streptavidin-HRP conjugate and a chemiluminescent substrate.

Luciferase reporter assays. MYC-335 elements containing G and T alleles of rs6983267 (1,406 bp) were amplified from genomic DNA of LoVo cells and cloned in front of a chicken delta 1-crystallin minimal promoter driving a firefly luciferase gene[47] (for PCR primers see **Supplementary Table 2**). The second and third nucleotides (AT) in the TCF4 sites were mutated to G and C using site-directed mutagenesis, resulting in mutant TCF4 sites GGCGAAAGG and GGCGAAAGT for the G and T allele, respectively. These mutations in TCF4 site completely abolish calculated binding affinity of TCF4 and the same nucleotides are also mutated in the FOPflash Wnt pathway negative control luciferase reporter plasmid. The constructs were fully sequenced and found to harbor three additional variant bases naturally occurring in LoVo, mapping outside the enhancer element. The presence of unintentionally introduced mutations was excluded. The reporters containing the 1,406-bp element were transfected into human HEK293T cells together with a *Renilla* luciferase control reporter. Where indicated, 20 mM LiCl was added after 24 h to induce Wnt pathway activity. Luciferase activities were measured at 32 h with the DualLuc kit (Promega). We calculated relative luciferase activities by dividing the firefly luciferase counts with the *Renilla* controls, and the results were normalized to the T allele without Wnt pathway induction. We controlled for the induction of Wnt pathway by LiCl by performing a parallel experiment using the SuperTOPflash Wnt pathway reporter. This reporter showed approximately 1,000-fold induction under these conditions (data not shown).

Chromatin immunoprecipitation (ChIP). We carried out ChIP essentially as described[48]. CRC cells were cross-linked with 1% formaldehyde in PBS at room temperature for 10 min, and reaction was stopped with 125 mM glycine.We collected the cells by centrifugation, isolated nuclei with hypotonic lysis buffer (1 mM EDTA, 10 mM KCl, 20 mM Hepes pH 7.9, 10% glycerol, 1 mM DTT with protease inhibitors) and lysed and sonicated cells in lysis buffer (0,1% SDS, 0,1% sodium deoxycholate, 1 mM EDTA, 10 mM TrisHCl pH 8.0, 140 mM NaCl, 1% Triton X-100 with protease inhibitors) to generate chromatin fragments of 100–400 bp in length. The fragmented chromatin material was clarified by centrifugation and pre-cleared with protein A-agarose beads, followed by immunoprecipitation with antibodies to TCF-4 (Mouse Anti-TCF-4 Monoclonal Antibody, Clone 6H5-3 Exalpha Biologicals), β-catenin (Rabbit Polyclonal Antibody: H-102, Santa-Cruz Biotechnology) or control IgG (normal mouse IgG: sc-2025 or normal rabbit IgG: sc-2027, Santa-Cruz Biotechnology). The precipitates were heat-treated to reverse the formaldehyde cross-links, followed by proteinase K and RNase A treatment, phenol-extraction and ethanol-precipitation. We assessed enrichment of immunoprecipitated material relative to the input material using PCR and gel electrophoresis, as well as quantitative real-time PCR. Allele-specific binding of TCF4 to the rs6983267-containing site was scored by comparing the height of the T and G peaks in sequencing tracts. This method was calibrated by PCR with different ratios of GG and TT DNAs and two distinct set of primers were used to confirm the results. Fold changes were calculated relative to IgG control. For primer sequences for the rs6983267 and negative and positive control regions, see **Supplementary Table 2**.

For ChIP-by-sequencing, the precipitated DNA was repaired using Klenow and T4 DNA polymerases and T4 polynucleotide kinase (MBI Fermentas, Latvia), and ligated to adapters according to manufacturer's instructions

doi:10.1038/ng.406

(Illumina). Subsequently, PCR-amplified fragments of approximately 130–150 bp were sequenced using Illumina Genome Analyzer (Uppsala University and University of Helsinki). Sequencing reads (30 bp or 36 bp) were mapped to the human genome (NCBI36) using Maq software by H. Li, version 0.6.5. Only high-quality reads that could be reliably mapped (mapping quality score at least 30) were accepted, resulting in a total of 13.1 and 7.3 million reads from TCF4 LoVo ChIP and IgG control samples, and 4.1 and 4.4 million reads from GP5D TCF4 ChIP and IgG control samples. Each read was then extended to a sequence of 130 bp, and height was determined at each position as the number of overlapping sequences. This analysis yields a maximum peak width of 250 bp for one occupied TCF4 site (120 bp in both directions from approximately 10 bp site). The positions with a height of 10 or more were defined as peaks. For each peak, the total number of sequences in the continuous region of four or more overlapping sequences was compared to the number of sequences in the same region in the IgG control. The probability of observing the difference between the sequence counts in the ChIP sample and IgG control by chance was estimated using Winflat program[49]. The program was originally developed for digital gene expression analysis and it can take into account the uncertainty associated to low sequence counts and the difference in the total amount of ChIP and IgG control reads. In total, we observed 558 TCF4 peaks in LoVo and 1694 peaks in GP5D with a height of ≥ 10 and $P < 0.05$.

Mouse embryo analyses. Regulatory element analyses in transgenic mouse embryos and *in situ* hybridizations were carried out essentially as described[18]. To create the *LacZ* reporter construct, a 1,193-bp genomic fragment including the rs6983267[G] containing regulatory element was PCR-amplified with Phu polymerase (Finnzymes) from mouse BAC clone and cloned into pTKPD vector (for primers, see **Supplementary Table 2**). A mutant construct with two mutated TCF4 sites (GATGAAAGG → GATGAGGGG, AATCAAAGG → AATCAGGGG) was generated with PCR mutagenesis. The constructs were sequenced to verify the sequence and orientation of the inserts. The primers used to generate the *Myc in situ* hybridization probe are described in **Supplementary Table 2**.

Microarray gene expression analysis. We exampled possible correlation of the rs6983267 genotype to *MYC* expression in lymphoblasts using publicly available exon array (see Accession codes section in Methods, HuEx-1_0-st-v2, Affymetrix) and genotype data from 57 CEU HapMap individuals[50,51]. Probe-level signals were background-corrected using the on chip antigenomic background probe group and normalized using sketch quantile normalization. Probeset and transcript cluster level signal intenstiy estimates were generated using PLIER. The extended level probeset grouping was used to construct transcript clusters. All preprocessing steps were carried out using Affymetrix Power Tools (APT). We excluded probesets that were not significantly different ($P > 0.05$) from the background signal in at least 50% of the samples within each group. This latter signal used the DABG value calculated by PLIER. Association between genotype and signal intensity levels of the probesets or transcript clusters was carried out using linear modelling implemented in Limma package in R[52]. The correlation between technical replicates was estimated and taken into account in the analyses. *P* values were produced with empirical Bayes statistics.

Gene expression profiles of 34 CRCs, created using Affymetrix HG-U133A[53], were used for studying the possible correlation between the genotype of the SNP and expression of *MYC*. The distribution of the genotypes among the samples was 11 with GG, 16 GT and 7 TT. The gene expression data are available in Gene Expression Omnibus (see Accession codes section in Methods). GCRMA pre-processing using CDF-files customized for Ensembl genome database was performed[54], after which probe set ENSG00000136997_at was used in the analysis. We examined the association between the SNP genotype and expression of *MYC* by linear regression, using R software.

Analysis of relationship between rs6983267 and *MYC* in normal colon. Frozen sections from normal colon tissue specimens were stained with hematoxylin and eosin to ascertain the presence of epithelial cells. The normal colonic epithelial cells were scraped from the fresh frozen tissue samples. Total RNA was extracted with Trizol reagent (Invitrogen) and purified and concentrated with RNeasy MinElute Cleanup colums (Qiagen). cDNA synthesis was performed with M-MLV enzyme (Promega). Relative expression of *MYC* mRNA in the samples was determined with TaqMan chemistry and ABI Prism 7500 sequence detection system. Assays for *MYC* (Assay ID: Hs00153408_m1) and endogenous control *PGK* (ID: Hs99999906_m1) were purchased from Applied Biosystems.

URLs. UCSC genome browser, http://genome.ucsc.edu/; VISTA enhancer browser, http://enhancer.lbl.gov/; Computational Methods for Locating and Analyzing Conserved Gene Regulatory DNA Elements, http://urn.fi/URN:ISBN:978-952-10-4353-6.

43. Aaltonen, L.A. *et al.* Incidence of hereditary nonpolyposis colorectal cancer and the feasibility of molecular screening for the disease. *N. Engl. J. Med.* **338**, 1481–1487 (1998).
44. Salovaara, R. *et al.* Population-based molecular detection of hereditary nonpolyposis colorectal cancer. *J. Clin. Oncol.* **18**, 2193–2200 (2000).
45. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* advance online publication, doi:10.1126/science.1162327 (14 May 2009).
46. Hallikas, O. & Taipale, J. High-throughput assay for determining specificity and affinity of protein-DNA binding interactions. *Nat. Protoc.* **1**, 215–222 (2006).
47. Taipale, J., Cooper, M.K., Maiti, T. & Beachy, P.A. Patched acts catalytically to suppress the activity of Smoothened. *Nature* **418**, 892–897 (2002).
48. Turunen, M.M., Dunlop, T.W., Carlberg, C. & Väisanen, S. Selective use of multiple vitamin D response elements underlies the 1α,25-dihydroxyvitamin D3-mediated negative regulation of the human CYP27B1 gene. *Nucleic Acids Res.* **35**, 2734–2747 (2007).
49. Audic, S. & Claverie, J.M. The significance of digital gene expression profiles. *Genome Res.* **7**, 986–995 (1997).
50. Kwan, T. *et al.* Genome-wide analysis of transcript isoform variation in humans. *Nat. Genet.* **40**, 225–231 (2008).
51. The International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
52. Smyth, G.K. in *Bioinformatics and Computional Biology Solutions using R and Bioconductor* (ed. Gentleman. R.) Limma:linear models for microarray data (Springer, New York, 2005).
53. Laiho, P. *et al.* Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene* **26**, 312–320 (2007).
54. Dai, M. *et al.* Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.* **33**, e175 (2005).