

Comprehensive molecular characterization of human colon and rectal cancer

The Cancer Genome Atlas Network*

To characterize somatic alterations in colorectal carcinoma, we conducted a genome-scale analysis of 276 samples, analysing exome sequence, DNA copy number, promoter methylation and messenger RNA and microRNA expression. A subset of these samples (97) underwent low-depth-of-coverage whole-genome sequencing. In total, 16% of colorectal carcinomas were found to be hypermutated: three-quarters of these had the expected high microsatellite instability, usually with hypermethylation and *MLH1* silencing, and one-quarter had somatic mismatch-repair gene and polymerase ϵ (*POLE*) mutations. Excluding the hypermutated cancers, colon and rectum cancers were found to have considerably similar patterns of genomic alteration. Twenty-four genes were significantly mutated, and in addition to the expected *APC*, *TP53*, *SMAD4*, *PIK3CA* and *KRAS* mutations, we found frequent mutations in *ARID1A*, *SOX9* and *FAM123B*. Recurrent copy-number alterations include potentially drug-targetable amplifications of *ERBB2* and newly discovered amplification of *IGF2*. Recurrent chromosomal translocations include the fusion of *NAV2* and WNT pathway member *TCF7L1*. Integrative analyses suggest new markers for aggressive colorectal carcinoma and an important role for *MYC*-directed transcriptional activation and repression.

The Cancer Genome Atlas project plans to profile genomic changes in 20 different cancer types and has so far published results on two cancer types^{1,2}. We now present results from multidimensional analyses of human colorectal carcinoma (CRC).

CRC is an important contributor to cancer mortality and morbidity. The distinction between the colon and the rectum is largely anatomical, but it has both surgical and radiotherapeutic management implications and it may have an impact on prognosis. Most investigators divide CRC biologically into those with microsatellite instability (MSI; located primarily in the right colon and frequently associated with the CpG island methylator phenotype (CIMP) and hyper-mutation) and those that are microsatellite stable but chromosomally unstable.

A rich history of investigations (for a review see ref. 3) has uncovered several critical genes and pathways important in the initiation and progression of CRC (ref. 3). These include the WNT, RAS–MAPK, PI3K, TGF- β , P53 and DNA mismatch-repair pathways. Large-scale sequencing analyses^{4–6} have identified numerous recurrently mutated genes and a recurrent chromosomal translocation. Despite this background, we have not had a fully integrated view of the genetic and genomic changes and their significance for colorectal tumorigenesis. Further insight into these changes may enable deeper understanding of the pathophysiology of CRC and may identify potential therapeutic targets.

Results

Tumour and normal pairs were analysed by different platforms. The specific numbers of samples analysed by each platform are shown in Supplementary Table 1.

Exome-sequence analysis

To define the mutational spectrum, we performed exome capture DNA sequencing on 224 tumour and normal pairs (all mutations are listed in Supplementary Table 2). Sequencing achieved >20-fold coverage of at least 80% of targeted exons. The somatic mutation rates varied considerably among the samples. Some had mutation rates of

<1 per 10⁶ bases, whereas a few had mutation rates of >100 per 10⁶. We separated cases (84%) with a mutation rate of <8.24 per 10⁶ (median number of non-silent mutations, 58) and those with mutation rates of >12 per 10⁶ (median number of total mutations, 728), which we designated as hypermutated (Fig. 1).

To assess the basis for the considerably different mutation rates, we evaluated MSI⁷ and mutations in the DNA mismatch-repair pathway^{8–10} genes *MLH1*, *MLH3*, *MSH2*, *MSH3*, *MSH6* and *PMS2*. Among the 30 hypermutated tumours with a complete data set, 23 (77%) had high levels of MSI (MSI-H). Included in this group were 19 tumours with *MLH1* methylation, 17 of which had CIMP. By comparison, the remaining seven hypermutated tumours, including the six with the highest mutation rates, lacked MSI-H, CIMP or *MLH1* methylation but usually had somatic mutations in one or more mismatch-repair genes or *POLE* aberrations seen rarely in the non-hypermutated tumours (Fig. 1).

Gene mutations

Overall, we identified 32 somatic recurrently mutated genes (defined by MutSig¹¹ and manual curation) in the hypermutated and non-hypermutated cancers (Fig. 1b). After removal of non-expressed genes, there were 15 and 17 in the hypermutated and non-hypermutated cancers, respectively (Fig. 1b; for a complete list see Supplementary Table 3). Among the non-hypermutated tumours, the eight most frequently mutated genes were *APC*, *TP53*, *KRAS*, *PIK3CA*, *FBXW7*, *SMAD4*, *TCF7L2* and *NRAS*. As expected, the mutated *KRAS* and *NRAS* genes usually had oncogenic codon 12 and 13 or codon 61 mutations, whereas the remaining genes had inactivating mutations. *CTNBN1*, *SMAD2*, *FAM123B* (also known as *WTX*) and *SOX9* were also mutated frequently. *FAM123B* is an X-linked negative regulator of WNT signalling¹², and virtually all of its mutations were loss of function. Mutations in *SOX9*, a gene important for cell differentiation in the intestinal stem cell niche^{13,14}, have not been associated previously with human cancer, but all nine mutated alleles in the non-hypermutated CRCs were frameshift or nonsense mutations. Tumour-suppressor

*Lists of participants and their affiliations appear at the end of the paper.

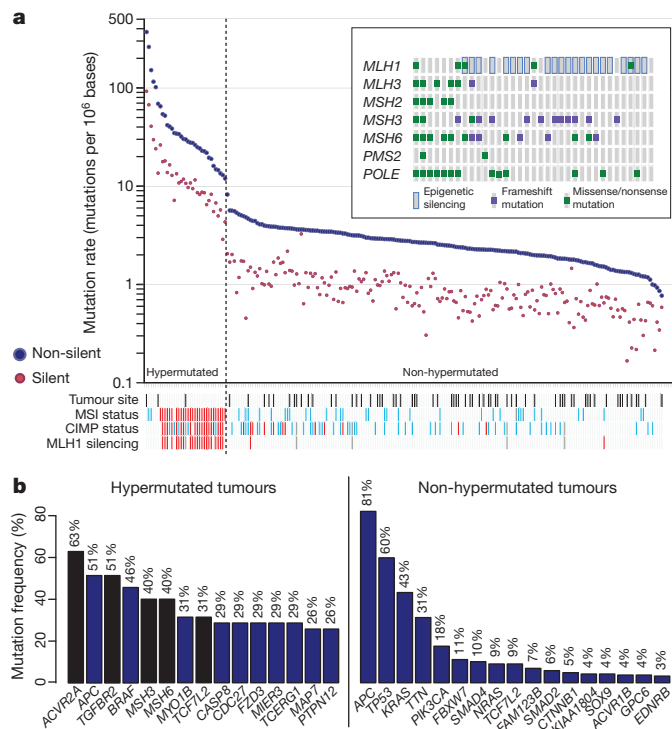


Figure 1 | Mutation frequencies in human CRC. **a**, Mutation frequencies in each of the tumour samples from 224 patients. Note a clear separation of hypermutated and non-hypermutated samples. Red, MSI high, CIMP high or MLH1 silenced; light blue, MSI low, or CIMP low; black, rectum; white, colon; grey, no data. Inset, mutations in mismatch-repair genes and *POLE* among the hypermutated samples. The order of the samples is the same as in the main graph. **b**, Significantly mutated genes in hypermutated and non-hypermutated tumours. Blue bars represent genes identified by the MutSig algorithm and black bars represent genes identified by manual examination of sequence data.

genes *ATM* and *ARID1A* also had a disproportionately high number of frameshift or nonsense mutations. *ARID1A* mutations have recently been reported in CRC and many other cancers^{15,16}.

In the hypermutated tumours, *ACVR2A*, *APC*, *TGFBR2*, *MSH3*, *MSH6*, *SLC9A9* and *TCF7L2* were frequent targets of mutation (Fig. 1b), along with mostly *BRAF(V600E)* mutations. However, two genes that were frequently mutated in the non-hypermutated

cancers were significantly less frequently mutated in hypermutated tumours: *TP53* (60 versus 20%, $P < 0.0001$) and *APC* (81% versus 51%, $P = 0.0023$; both Fisher's exact test). Other genes, including *TGFBR2*, were mutated recurrently in the hypermutated cancers, but not in the non-hypermutated samples. These findings indicate that hypermutated and non-hypermutated tumours progress through different sequences of genetic events.

As expected, hypermutated tumours with *MLH1* silencing and MSI-H showed additional differences in the mutational profile. When we specifically examined 28 genes with long mononucleotide repeats in their coding sequences, we found that the rate of frameshift mutation was 3.6-fold higher than the rate of such mutations in hypermutated tumours without *MLH1* silencing and 50-fold higher than that in non-hypermutated tumours (Supplementary Table 2).

Mutation rate and methylation patterns

As mentioned above, patients with colon and rectal tumours are managed differently¹⁷, and epidemiology also highlights differences between the two¹⁷. An initial integrative analysis of MSI status, somatic copy-number alterations (SCNAs), CIMP status and gene-expression profiles of 132 colonic and 62 rectal tumours enabled us to examine possible biological differences between tumours in the two locations. Among the non-hypermutated tumours, however, the overall patterns of changes in copy number, CIMP, mRNA and miRNA were indistinguishable between colon and rectal carcinomas (Fig. 2). On the basis of this result, we merged the two for all subsequent analyses.

Unsupervised clustering of the promoter DNA methylation profiles of 236 colorectal tumours identified four subgroups (Supplementary Fig. 1 and Supplementary Methods). Two of the clusters contained tumours with elevated rates of methylation and were classified as CIMP high and CIMP low, as previously described¹⁸. The two non-CIMP clusters were predominantly from tumours that were non-hypermutated and derived from different anatomic locations. mRNA expression profiles separated the colorectal tumours into three distinct clusters (Supplementary Fig. 2). One significantly overlapped with CIMP-high tumours ($P = 3 \times 10^{-12}$) and was enriched with hypermutated tumours, and the other two clusters did not correspond with any group in the methylation data. Analysis of miRNA expression by unsupervised clustering (Supplementary Fig. 3) identified no clear distinctions between rectal cancers and non-hypermutated colon cancers.

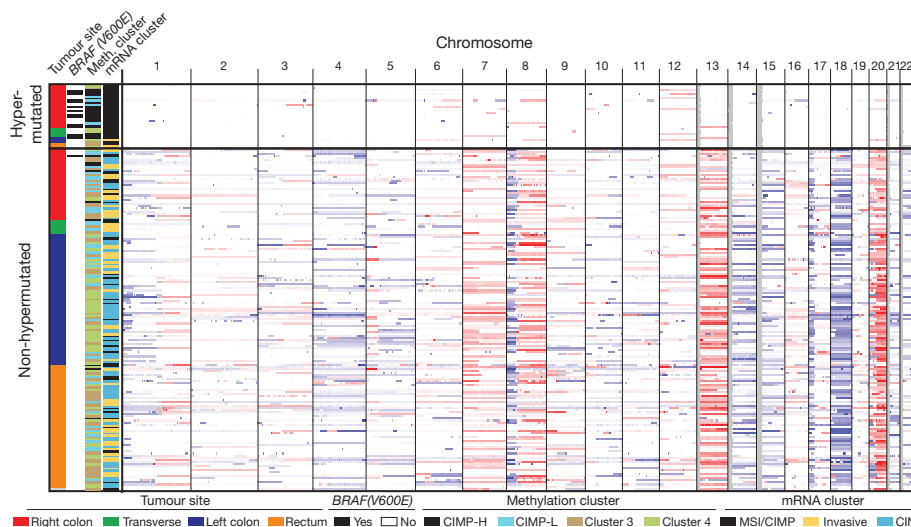


Figure 2 | Integrative analysis of genomic changes in 195 CRCs. Hypermutated tumours have near-diploid genomes and are highly enriched for hypermethylation, CIMP expression phenotype and *BRAF(V600E)* mutations. Non-hypermutated tumours originating from different sites are virtually

indistinguishable from each other on the basis of their copy-number alteration patterns, DNA methylation or gene-expression patterns. Copy-number changes of the 22 autosomes are shown in shades of red for copy-number gains and shades of blue for copy-number losses.

Chromosomal and sub-chromosomal changes

In total, 257 tumours were profiled for SCNAs with Affymetrix SNP 6.0 arrays. Of these tumours, 97 were also analysed by low-depth-of-coverage (low-pass) whole-genome sequencing. As expected, the hypermutated tumours had far fewer SCNAs (Fig. 2). No difference was found between microsatellite-stable and -unstable hypermutated tumours (Supplementary Fig. 4). We used the GISTIC algorithm¹⁹ to identify probable gene targets of focal alterations. There were several previously well-defined arm-level changes, including gains of 1q, 7p and q, 8p and q, 12q, 13q, 19q, and 20p and q (ref. 6). (Supplementary Fig. 4 and Supplementary Table 4). Significantly deleted chromosome arms were 18p and q (including *SMAD4*) in 66% of the tumours and 17p and q (including *TP53*) in 56%. Other significantly deleted chromosome arms were 1p, 4q, 5q, 8p, 14q, 15q, 20p and 22q.

We identified 28 recurrent deletion peaks (Supplementary Fig. 4 and Supplementary Table 4), including the genes *FHIT*, *RBFOX1* and *WWOX* with large genomic footprints located in potentially fragile sites of the genome, in near-diploid hypermutated tumours. Other focal deletions involved tumour-suppressor genes such as *SMAD4*, *APC*, *PTEN* and *SMAD3*. A significant focal deletion of 10p25.2 spanned four genes, including *TCF7L2*, which was also frequently mutated in our data set. A gene fusion between adjacent genes *VTT1A* and *TCF7L2* through an interstitial deletion was found in 3% of CRCs and is required for survival of CRC cells bearing the translocation⁴.

There were 17 regions of significant focal amplification (Supplementary Table 4). Some of these were superimposed on broad gains of chromosome arms, and included a peak at 13q12.13 near the peptidase-coding gene *USP12* and at ~500 kb distal to the CRC

candidate oncogene *CDK8*; an adjacent peak at 13q12; a peak containing *KLF5* at 13q22.1; and a peak at 20q13.12 adjacent to *HNF4A*. Peaks on chromosome 8 included 8p12 (which contains the histone methyl-transferase-coding gene *WHSC1L1*, adjacent to *FGFR1*) and 8q24 (which contains *MYC*). An amplicon at 17q21.1, found in 4% of the tumours, contains seven genes, including the tyrosine kinase *ERBB2*. *ERBB2* amplifications have been described in colon, breast and gastro-oesophageal tumours, and breast and gastric cancers bearing these amplifications have been treated effectively with the anti-*ERBB2* antibody trastuzumab^{20–22}.

One of the most common focal amplifications, found in 7% of the tumours, is the gain of a 100–150-kb region of the chromosome arm 11p15.5. It contains genes encoding insulin (*INS*), insulin-like growth factor 2 (*IGF2*) and tyrosine hydroxylase (*TH*), as well as *miR-483*, which is embedded within *IGF2* (Fig. 3a). We found elevated expression of *IGF2* and *miR-483* but not of *INS* and *TH* (Fig. 3b, c). Immediately adjacent to the amplified region is *ASCL2*, a transcription factor active in specifying intestinal stem-cell fate²³. Although *ASCL2* has been implicated as a target of amplification in CRC^{23–25}, it was consistently outside the region of amplification and its expression was not correlated with copy-number changes. These observations suggest that *IGF2* and *miR-483* are candidate functional targets of 11p15.5 amplification. *IGF2* overexpression through loss of imprinting has been implicated in the promotion of CRC^{26, 27}. *MiR-483* may also have a role in CRC pathogenesis²⁸.

A subset of tumours without *IGF2* amplification (15%) also had considerably higher levels of *IGF2* gene expression (as much as a 100-fold increase), an effect not attributable to methylation changes at the *IGF2* promoter. To assess the context of *IGF2* amplification/

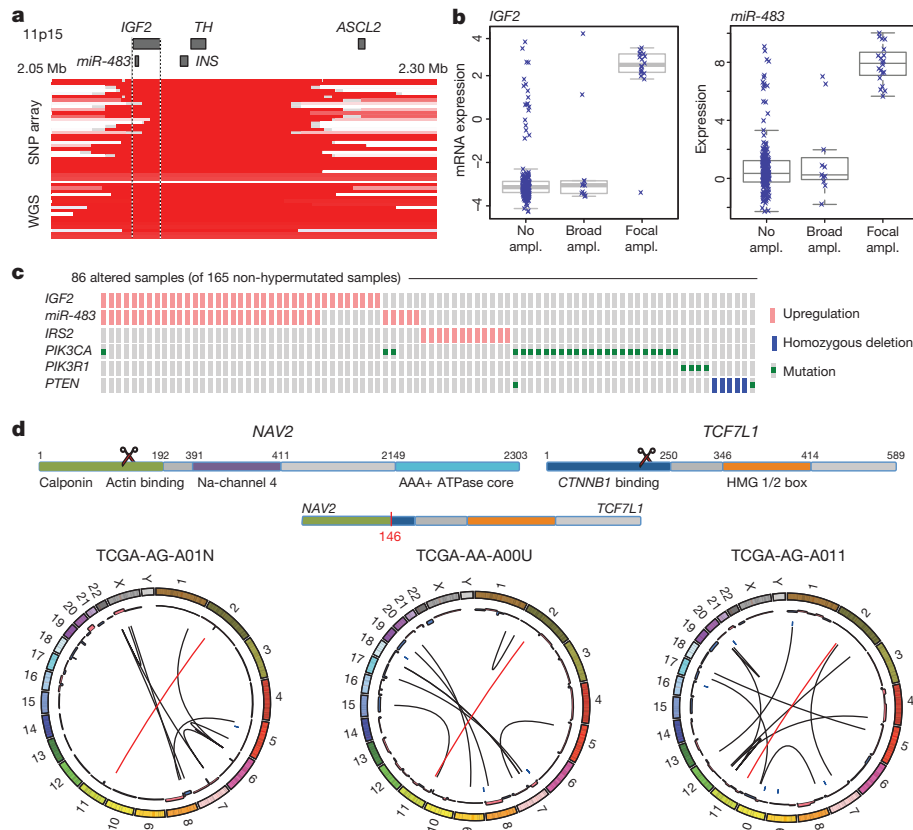


Figure 3 | Copy-number changes and structural aberrations in CRC.

a, Focal amplification of 11p15.5. Segmented DNA copy-number data from single-nucleotide polymorphism (SNP) arrays and low-pass whole-genome sequencing (WGS) are shown. Each row represents a patient; amplified regions are shown in red. **b**, Correlation of expression levels with copy-number changes for *IGF2* and *miR-483*. **c**, *IGF2* amplification and overexpression are mutually

exclusive of alterations in PI3K signalling-related genes. **d**, Recurrent *NAV2-TCF7L2* fusions. The structure of the two genes, locations of the breakpoints leading to the translocation and circular representations of all rearrangements in tumours with a fusion are shown. Red lines represent the *NAV2-TCF7L2* fusions and black lines represent other rearrangements. The inner ring represents copy-number changes (blue denotes loss, pink denotes gain).

overexpression, we systematically searched for mutually exclusive genomic events using the MEMo method²⁹. We found a pattern of near exclusivity (corrected $P < 0.01$) of *IGF2* overexpression with genomic events known to activate the PI3K pathway (mutations of *PIK3CA* and *PIK3R1* or deletion/mutation of *PTEN*; Fig. 3c and Supplementary Table 5). The *IRS2* gene, encoding a protein linking IGF1R (the receptor for IGF2) with PI3K, is on chromosome 13, which is frequently gained in CRC. The cases with the highest *IRS2* expression were mutually exclusive of the cases with *IGF2* overexpression ($P = 0.04$) and also lacked mutations in the PI3K pathway ($P = 0.0001$; Fig. 3c). These results strongly suggest that the IGF2–IGF1R–*IRS2* axis signals to PI3K in CRC and imply that therapeutic targeting of the pathway could act to block PI3K activity in this subset of patients.

Translocations

To identify new chromosomal translocations, we performed low-pass, paired-end, whole-genome sequencing on 97 tumours with matched normal samples. In each case we achieved sequence coverage of ~3–4-fold and a corresponding physical coverage of 7.5–10-fold. Despite the low genome coverage, we detected 250 candidate interchromosomal translocation events (range, 0–10 per tumour). Among these events, 212 had one or both breakpoints in an intergenic region, whereas the remaining 38 juxtaposed coding regions of two genes in putative fusion events, of which 18 were predicted to code for in-frame events (Supplementary Table 6). We found three separate cases in which the first two exons of the *NAV2* gene on chromosome 11 are joined with the 3' coding portion of *TCF7L1* on chromosome 2 (Supplementary Fig. 5). *TCF7L1* encodes TCF3, a member of the TCF/LEF class of transcription factors that heterodimerize with nuclear β -catenin to enable β -catenin-mediated transcriptional regulation. Intriguingly, in all three cases, the predicted structure of the *NAV2*–*TCF7L1* fusion protein lacks the TCF3 β -catenin-binding domain. This translocation is similar to another recurrent translocation identified in CRC, a fusion in which the amino terminus of *VTI1A* is joined to *TCF4*, which is encoded by *TCF7L2*, a homologue of *TCF7L1* that is deleted or mutated in 12% of non-hypermuted tumours⁴. We also observed 21 cases of translocation involving *TTC28* located on chromosome 22 (Supplementary Table 6). In all

cases the fusions predict inactivation of *TTC28*, which has been identified as a target of P53 and an inhibitor of tumour cell growth³⁰. Eleven of the 19 (58%) gene–gene translocations were validated by obtaining PCR products or, in some cases, sequencing the junction fragments (Supplementary Fig. 5).

Altered pathways in CRC

Integrated analysis of mutations, copy number and mRNA expression changes in 195 tumours with complete data enriched our understanding of how some well-defined pathways are deregulated. We grouped samples by hypermutation status and identified recurrent alterations in the WNT, MAPK, PI3K, TGF- β and p53 pathways (Fig. 4, Supplementary Fig. 6 and Supplementary Table 1).

We found that the WNT signalling pathway was altered in 93% of all tumours, including biallelic inactivation of *APC* (Supplementary Table 7) or activating mutations of *CTNNB1* in ~80% of cases. There were also mutations in *SOX9* and mutations and deletions in *TCF7L2*, as well as the DKK family members and *AXIN2*, *FBXW7* (Supplementary Fig. 7), *ARID1A* and *FAM123B* (the latter is a negative regulator of WNT– β -catenin signalling¹² found mutated in Wilms' tumour³¹). A few mutations in *FAM123B* have previously been described in CRC³². *SOX9* has been suggested to have a role in cancer, but no mutations have previously been described. The WNT receptor frizzled (*FZD10*) was overexpressed in ~17% of samples, in some instances at levels of 100 \times normal. Altogether, we found 16 different altered WNT pathway genes, confirming the importance of this pathway in CRC. Interestingly, many of these alterations were found in tumours that harbour *APC* mutations, suggesting that multiple lesions affecting the WNT signalling pathway confer selective advantage.

Genetic alterations in the PI3K and RAS–MAPK pathways are common in CRC. In addition to *IGF2* and *IRS2* overexpression, we found mutually exclusive mutations in *PIK3R1* and *PIK3CA* as well as deletions in *PTEN* in 2%, 15% and 4% of non-hypermuted tumours, respectively. We found that 55% of non-hypermuted tumours have alterations in *KRAS*, *NRAS* or *BRAF*, with a significant pattern of mutual exclusivity (Supplementary Fig. 6 and Supplementary Table 1). We also evaluated mutations in the erythroblastic leukemia viral oncogene homolog (ERBB) family of receptors because of the translational relevance of such mutations. Mutations or amplifications in

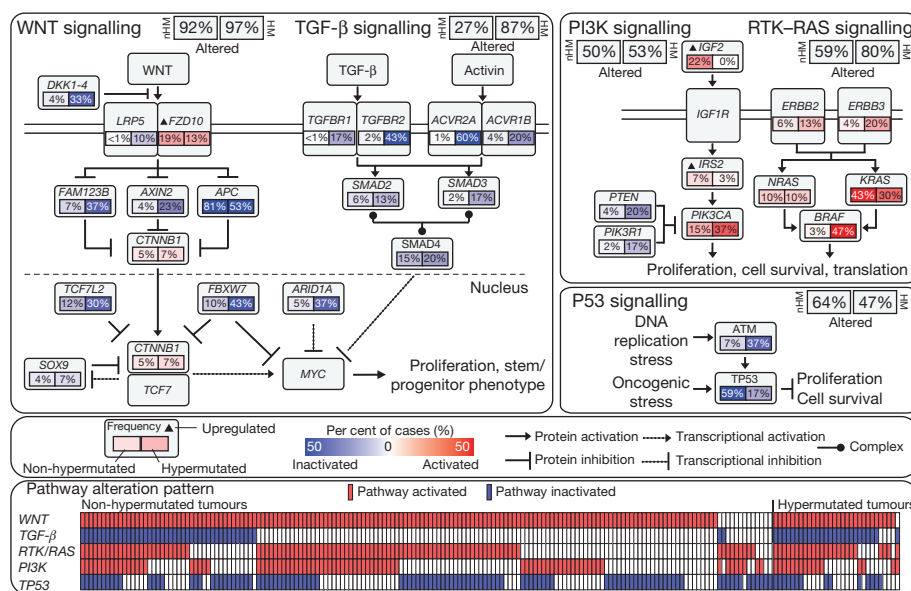


Figure 4 | Diversity and frequency of genetic changes leading to deregulation of signalling pathways in CRC. Non-hypermuted (nHM; $n = 165$) and hypermutated (HM; $n = 30$) samples with complete data were analysed separately. Alterations are defined by somatic mutations, homozygous deletions, high-level focal amplifications, and, in some cases, by significant

up- or downregulation of gene expression (*IGF2*, *FZD10*, *SMAD4*). Alteration frequencies are expressed as a percentage of all cases. Red denotes activated genes and blue denotes inactivated genes. Bottom panel shows for each sample if at least one gene in each of the five pathways described in this figure is altered.

one of the four ERBB family genes are present in 22 out of 165 (13%) non-hypermuted and 16 out of 30 (53%) hypermutated cases. Some of the mutations are listed in the COSMIC database³³, suggesting a functional role. Intriguingly, recurrent *ERBB2(V842I)* and *ERBB3(V104M)* mutations were found in four and two non-hypermuted cases, respectively. Mutations and focal amplifications of *ERBB2* (Supplementary Fig. 6) should be evaluated as predictors of response to agents that target those receptors. We observed co-occurrence of alterations involving the RAS and PI3K pathways in one-third of tumours (Fig. 4; $P = 0.039$, Fisher's exact test). These results indicate that simultaneous inhibition of the RAS and PI3K pathways may be required to achieve therapeutic benefit.

The TGF- β signalling pathway is known to be deregulated in CRC and other cancers³⁴. We found genomic alterations in *TGFBR1*, *TGFBR2*, *ACVR2A*, *ACVR1B*, *SMAD2*, *SMAD3* and *SMAD4* in 27% of the non-hypermuted and 87% of the hypermutated tumours. We also evaluated the p53 pathway, finding alterations in *TP53* in 59% of non-hypermuted cases (mostly biallelic; Supplementary Table 8) and alterations in *ATM*, a kinase that phosphorylates and activates P53 after DNA damage, in 7%. Alterations in these two genes showed a trend towards mutual exclusivity ($P = 0.016$) (Fig. 4, Supplementary Fig. 6 and Supplementary Table 1).

We integrated copy number, gene expression, methylation and pathway data using the PARADIGM software platform³⁵. The analysis showed a number of new characteristics of CRC (Fig. 5a). For example, despite the diversity in anatomical origin or mutation levels, nearly 100% of these tumours have changes in *MYC* transcriptional targets, both those promoted by and those inhibited by *MYC*. These findings are consistent with patterns deduced from genetic alterations (Fig. 4) and suggest an important role for *MYC* in CRC. The analysis also identified several gene networks altered across all tumour samples and those with differential alterations in hypermutated versus non-hypermuted samples (Supplementary Table 7, Supplementary Data on the Cancer Genome Atlas publication webpage).

Because most of the tumours used in this study were derived from a prospective collection, survival data are not available. However, the tumours can be classified as aggressive or non-aggressive on the basis

of tumour stage, lymph node status, distant metastasis and vascular invasion at the time of surgery. We found numerous molecular signatures associated with tumour aggressiveness, a subset of which is shown in Fig. 5b. They include specific focal amplifications and deletions, and altered gene-expression levels, including those of *SCN5A* (ref. 36), a reported regulator of colon cancer invasion (see Supplementary Tables 10 and 11 for a full list). Association with tumour aggressiveness is also observed in altered expression of miRNAs and specific somatic mutations (*APC*, *TP53*, *PIK3CA*, *BRAF* and *FBXW7*; Supplementary Fig. 8b). Mutations in *FBXW7* (38 cases) and distant metastasis (32 cases) never co-occurred ($P = 0.0019$). Interestingly, a number of genomic regions have multiple molecular associations with tumour aggressiveness that manifest as clinically related genomic hotspots. Examples of this are the region 20q13.12, which includes a focal amplification and multiple genes correlating with tumour aggression, and the region 22q12.3, containing *APOL6* (ref. 37) (Supplementary Figures 8 and 9).

Discussion

This comprehensive integrative analysis of 224 colorectal tumour and normal pairs provides a number of insights into the biology of CRC and identifies potential therapeutic targets. To identify possible biological differences in colon and rectum tumours, we found, in the non-hypermuted tumours irrespective of their anatomical origin, the same type of copy number, expression profile, DNA methylation and miRNA changes. Over 94% had a mutation in one or more members of the WNT signalling pathway, predominantly in *APC*. However, there were some differences between tumours from the right colon and all other sites. Hypermethylation was more common in the right colon, and three-quarters of hypermutated samples came from the same site, although not all of them had MSI (Fig. 2). Why most of the hypermutated samples came from the right colon and why there are two classes of tumours at this site is not known. The origins of the colon from embryonic midgut and hindgut may provide an explanation. As the survival rate of patients with high MSI-related cancers is better and these cancers are hypermutated, mutation rate may be a better prognostic indicator.

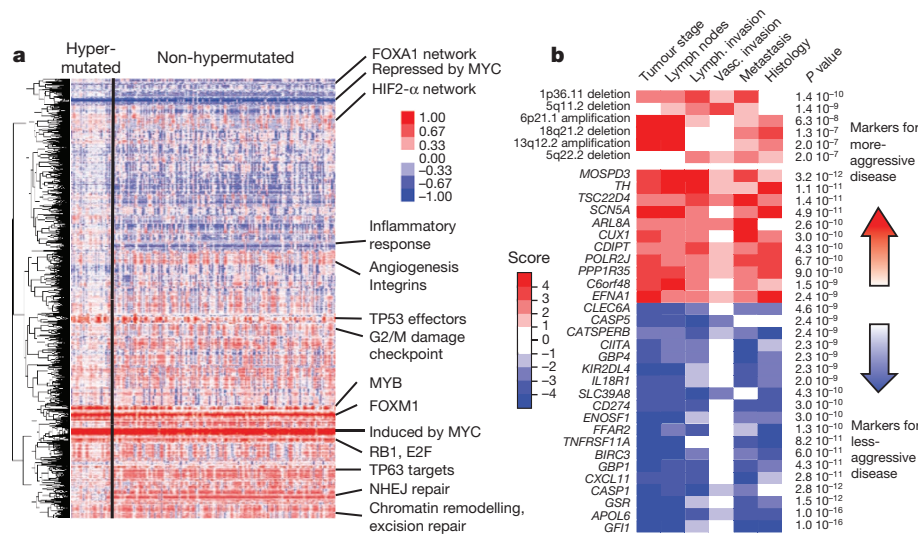


Figure 5 | Integrative analyses of multiple data sets. **a**, Clustering of genes and pathways affected in colon and rectum tumours deduced by PARADIGM analysis. Blue denotes under-expressed relative to normal and red denotes overexpressed relative to normal. Some of the pathways deduced by this method are shown on the right. NHEJ, non-homologous end joining. **b**, Gene-expression signatures and SCNAs associated with tumour aggressiveness. Molecular signatures (rows) that show a statistically significant association with tumour aggressiveness according to selected clinical assays (columns) are shown in colour, with red indicating markers of tumour aggressiveness and

blue indicating the markers of less-aggressive tumours. Significance is based on the combined P value from the weighted Fisher's method, corrected for multiple testing. Colour intensity and score is in accordance with the strength of an individual clinical-molecular association, and is proportional to $\log_{10}(P)$, where P is the P value for that association. To limit the vertical extent of the figure, gene-expression signatures are restricted to a combined P value of $P < 10^{-9}$ and SCNAs to $P < 10^{-7}$, and features are shown only if they are also significant in the subset of non-MSI-H samples (the analysis was performed separately on the full data as well as on the MSI-H and non-MSI-H subgroups).

Whole-exome sequencing and integrative analysis of genomic data provided further insights into the pathways that are dysregulated in CRC. We found that 93% of non-hypermethylated and 97% of hypermethylated cases had a deregulated WNT signalling pathway. New findings included recurrent mutations in *FAM123B*, *ARID1A* and *SOX9* and very high levels of overexpression of the WNT ligand receptor gene *FZD10*. To our knowledge, *SOX9* has not previously been described as frequently mutated in any human cancer. *SOX9* is transcriptionally repressed by WNT signalling, and the *SOX9* protein has been shown to facilitate β -catenin degradation³⁸. *ARID1A* is frequently mutated in gynaecological cancers and has been shown to suppress *MYC* transcription³⁹. Activation of WNT signalling and inactivation of the TGF- β signalling pathway are known to result in activation of *MYC*. Our mutational and integrative analyses emphasize the critical role of *MYC* in CRC. We also compared our results with other large-scale analyses⁶ and found many similarities and few differences in mutated genes (Supplementary Table 3).

Our integrated analysis revealed a diverse set of changes in TCF/LEF-encoding genes, suggesting additional roles for TCF/LEF factors in CRC beyond being passive partners for β -catenin.

Our data suggest a number of therapeutic approaches to CRC. Included are WNT-signalling inhibitors and small-molecule β -catenin inhibitors, which are showing initial promise^{40–42}. We find that several proteins in the RTK-RAS and PI3K pathways, including IGF2, IGFR, ERBB2, ERBB3, MEK, AKT and MTOR could be targets for inhibition.

Our analyses show that non-hypermethylated adenocarcinomas of the colon and rectum are not distinguishable at the genomic level. However, tumours from the right/ascending colon were more likely to be hypermethylated and to have elevated mutation rates than were other CRCs. As has been recognized previously, activation of the WNT signalling pathway and inactivation of the TGF- β signalling pathway, resulting in increased activity of *MYC*, are nearly ubiquitous events in CRC. Genomic aberrations frequently target the MAPK and PI3K pathways but less frequently target receptor tyrosine kinases. In conclusion, the data presented here provide a useful resource for understanding this deadly disease and identifying possibilities for treating it in a targeted way.

METHODS SUMMARY

Tumour and normal samples were processed by either of two biospecimen core resources, and aliquots of purified nucleic acids were shipped to the genome characterization and sequencing centres (Supplementary Methods). The biospecimen core resources provided sample sets in several different batches. To assess any batch effects we examined the mRNA expression, miRNA expression and DNA methylation data sets using a combination of cluster analysis, enhanced principal component analysis and analysis of variance (Supplementary Methods). Although some differences among batches were detected, we did not correct them computationally because the differences were generally modest and because some of them may reflect biological phenomena (Supplementary Methods).

We used Affymetrix SNP 6.0 microarrays to detect copy-number alterations. A subset of samples was subjected to low-pass (2–5 \times) whole-genome sequencing (Illumina HiSeq), in part for detection of SCNA and chromosomal translocations^{43,44}. Gene-expression profiles were generated using Agilent microarrays and RNA-Seq. DNA methylation data were obtained using Illumina Infinium (HumanMethylation27) arrays. DNA sequencing of coding regions was performed by exome capture followed by sequencing on the SOLiD or Illumina HiSeq platforms. Details of the analytical methods used are described in Supplementary Methods.

All of the primary sequence files are deposited in dbGap and all other data are deposited at the Data Coordinating Center (DCC) for public access (<http://cancergenome.nih.gov/>). Data matrices and supporting data can be found at http://tcga-data.nci.nih.gov/docs/publications/coadread_2012/. The data can also be explored through the ISB Regulome Explorer (<http://explorer.cancerregulome.org/>), Next Generation Clustered Heat Maps (<http://bioinformatics.mdanderson.org/main/TCGA/Supplements/NGCHM-CRC>) and the cBio Cancer Genomics Portal (<http://cbioportal.org>). Descriptions of the data can be found at <https://wiki.nci.nih.gov/x/j5dXAg> and in Supplementary Methods.

Received 15 November 2011; accepted 22 May 2012.

1. The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**, 1061–1068 (2008).
2. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
3. Fearon, E. R. Molecular genetics of colorectal cancer. *Annu. Rev. Pathol.* **6**, 479–507 (2011).
4. Bass, A. J. *et al.* Genomic sequencing of colorectal adenocarcinomas identifies a recurrent *VTI1A-TCF7L2* fusion. *Nature Genet.* **43**, 964–968 (2011).
5. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
6. Wood, L. D. *et al.* The genomic landscapes of human breast and colorectal cancers. *Science* **318**, 1108–1113 (2007).
7. Umar, A. *et al.* Revised Bethesda guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *J. Natl. Cancer Inst.* **96**, 261–268 (2004).
8. Aaltonen, L. A. *et al.* Clues to the pathogenesis of familial colorectal cancer. *Science* **260**, 812–816 (1993).
9. Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D. & Perucho, M. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature* **363**, 558–561 (1993).
10. Parsons, R. *et al.* Hypermethylability and mismatch repair deficiency in RER⁺ tumor cells. *Cell* **75**, 1227–1236 (1993).
11. Dooley, A. L. *et al.* Nuclear factor I/B is an oncogene in small cell lung cancer. *Genes Dev.* **25**, 1470–1475 (2011).
12. Major, M. B. *et al.* Wilms tumor suppressor *WTX* negatively regulates WNT/ β -catenin signaling. *Science* **316**, 1043–1046 (2007).
13. Mori-Akiyama, Y. *et al.* *SOX9* is required for the differentiation of paneth cells in the intestinal epithelium. *Gastroenterology* **133**, 539–546 (2007).
14. Bastide, P. *et al.* *Sox9* regulates cell proliferation and is required for Paneth cell differentiation in the intestinal epithelium. *J. Cell Biol.* **178**, 635–648 (2007).
15. Jones, S. *et al.* Somatic mutations in the chromatin remodeling gene *ARID1A* occur in several tumor types. *Hum. Mutat.* **33**, 100–103 (2012).
16. Wilson, B. G. & Roberts, C. W. *SWI/SNF* nucleosome remodellers and cancer. *Nat. Rev. Cancer* **11**, 481–492 (2011).
17. Minsky, B. D. Unique considerations in the patient with rectal cancer. *Semin. Oncol.* **38**, 542–551 (2011).
18. Hinoue, T. *et al.* Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Res.* **22**, 271–282 (2012).
19. Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proc. Natl. Acad. Sci. USA* **104**, 20007–20012 (2007).
20. Camps, J. *et al.* Integrative genomics reveals mechanisms of copy number alterations responsible for transcriptional deregulation in colorectal cancer. *Genes Chromosom. Cancer* **48**, 1002–1017 (2009).
21. Varley, J. M., Swallow, J. E., Brammar, W. J., Whittaker, J. L. & Walker, R. A. Alterations to either *c-erbB-2*(neu) or *c-myc* proto-oncogenes in breast carcinomas correlate with poor short-term prognosis. *Oncogene* **1**, 423–430 (1987).
22. Yokota, J. *et al.* Amplification of *c-erbB-2* oncogene in human adenocarcinomas *in vivo*. *Lancet* **327**, 765–767 (1986).
23. van der Flier, L. G. *et al.* Transcription factor achaete scute-like 2 controls intestinal stem cell fate. *Cell* **136**, 903–912 (2009).
24. Jubb, A. M., Hoefflich, K. P., Haverty, P. M., Wang, J. & Koeppen, H. *Ascl2* and *11p15.5* amplification in colorectal cancer. *Gut* **60**, 1606–1607 (2011).
25. Stange, D. E. *et al.* Expression of an *ASCL2* related stem cell signature and *IGF2* in colorectal cancer liver metastases with *11p15.5* gain. *Gut* **59**, 1236–1244 (2010).
26. Cui, H. *et al.* Loss of *IGF2* imprinting: a potential marker of colorectal cancer risk. *Science* **299**, 1753–1755 (2003).
27. Nakagawa, H. *et al.* Loss of imprinting of the insulin-like growth factor II gene occurs by allelic methylation in a core region of *H19*-associated CTCF-binding sites in colorectal cancer. *Proc. Natl. Acad. Sci. USA* **98**, 591–596 (2001).
28. Veronese, A. *et al.* Oncogenic role of *miR-483-3p* at the *IGF2/483* locus. *Cancer Res.* **70**, 3140–3149 (2010).
29. Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.* **22**, 398–406 (2012).
30. Brady, C. A. *et al.* Distinct p53 transcriptional programs dictate acute DNA-damage responses and tumor suppression. *Cell* **145**, 571–583 (2011).
31. Rivera, M. N. *et al.* An X chromosome gene, *WTX*, is commonly inactivated in Wilms tumor. *Science* **315**, 642–645 (2007).
32. Scheel, S. K. *et al.* Mutations in the *WTX*-gene are found in some high-grade microsatellite instable (MSI-H) colorectal cancers. *BMC Cancer* **10**, 413 (2010).
33. Forbes, S. A. *et al.* The catalogue of somatic mutations in cancer (COSMIC). *Curr. Protoc. Hum. Genet.* Ch. 10, Unit 10.11 (2008).
34. Massagué, J., Blain, S. W. & Lo, R. S. TGF β signaling in growth control, cancer, and heritable disorders. *Cell* **103**, 295–309 (2000).
35. Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics* **26**, i237–i245 (2010).
36. House, C. D. *et al.* Voltage-gated Na⁺ channel *SCN5A* is a key regulator of a gene transcriptional network that controls colon cancer invasion. *Cancer Res.* **70**, 6957–6967 (2010).
37. Liu, Z., Lu, H., Jiang, Z., Pastuszyn, A. & Hu, C. A. Apolipoprotein I6, a novel proapoptotic Bcl-2 homology 3-only protein, induces mitochondria-mediated apoptosis in cancer cells. *Mol. Cancer Res.* **3**, 21–31 (2005).

38. Topol, L., Chen, W., Song, H., Day, T. F. & Yang, Y. Sox9 inhibits Wnt signaling by promoting β -catenin phosphorylation in the nucleus. *J. Biol. Chem.* **284**, 3323–3333 (2009).
39. Nagl, N. G. Jr, Zweitzig, D. R., Thimmapaya, B., Beck, G. R. Jr & Moran, E. The *c-myc* gene is a direct target of mammalian SWI/SNF-related complexes during differentiation-associated cell cycle arrest. *Cancer Res.* **66**, 1289–1293 (2006).
40. Chen, B. *et al.* Small molecule-mediated disruption of Wnt-dependent signaling in tissue regeneration and cancer. *Nat. Chem. Biol.* **5**, 100–107 (2009).
41. Ewan, K. *et al.* A useful approach to identify novel small-molecule inhibitors of Wnt-dependent transcription. *Cancer Res.* **70**, 5963–5973 (2010).
42. Sack, U. *et al.* S100A4-induced cell motility and metastasis is restricted by the Wnt/ β -catenin pathway inhibitor calcimycin in colon cancer cells. *Mol. Biol. Cell* **22**, 3344–3354 (2011).
43. Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature Methods* **6**, 677–681 (2009).
44. Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc. Natl Acad. Sci. USA* **108**, E1128–E1136 (2011).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by the following grants from the National Institutes of Health: U24CA143799, U24CA143835, U24CA143840, U24CA143843, U24CA143845, U24CA143848, U24CA143858, U24CA143866, U24CA143867, U24CA143882, U24CA143883, U24CA144025, U54HG003067, U54HG003079 and U54HG003273.

Author Contributions The Cancer Genome Atlas research network contributed collectively to this study. Biospecimens were provided by the tissue source sites and processed by the Biospecimen Core Resource. Data generation and analyses were performed by the genome-sequencing centers, cancer genome-characterization centers and genome data analysis centers. All data were released through the Data Coordinating Center. Project activities were coordinated by the National Cancer Institute and National Human Genome Research Institute project teams. Project leaders were R.K. and D.A.W. Writing team, T.A., A.J.B., T.A.C., L.D., A.H., S.R.H., R.K., P.W.L., M.M., N.S., I.S., J.M.S., J.T., V.T. and D.A.W.; mutations, M.S.L., L.R.T., D.A.W. and G.G.; copy-number and structural aberrations, A.H.R., A.J.B., A.H. and P.-C.C.; DNA methylation, T.H.; expression, J.T.A.; miRNA, G.R., A.C.; pathways, C.J.C., L.D., T.G., S.N., J.D.R., C.S., N.S., J.M.S. and V.T.

Author Information dbGaP accession numbers have been provided in Supplementary Table 1. The authors declare no competing financial interests. Reprints and permissions information is available at www.nature.com/reprints. Readers are welcome to comment on the online version of this article at www.nature.com/nature. This paper is distributed under the terms of the Creative Commons Attribution-Non-Commercial-Share Alike licence, and is freely available to all readers at www.nature.com/nature. Correspondence and requests for materials should be addressed to R.K. (rkucherlapati@partners.org).

Genomic Sequencing Center Baylor College of Medicine Donna M. Muzny¹, Matthew N. Bainbridge¹, Kyle Chang¹, Huyen H. Dinh¹, Jennifer A. Drummond¹, Gerald Fowler¹, Christie L. Kovar¹, Lora R. Lewis¹, Margaret B. Morgan¹, Irene F. Newsham¹, Jeffrey G. Reid¹, Jireh Santibanez¹, Eve Shinbrot¹, Lisa R. Trevino¹, Yuan-Qing Wu¹, Min Wang¹, Preethi Gunaratne^{1,2}, Lawrence A. Donehower^{1,3}, Chad J. Creighton^{1,3}, David A. Wheeler¹, Richard A. Gibbs¹; **Genome Sequencing Center Broad Institute** Michael S. Lawrence⁴, Douglas Voet⁴, Rui Jing⁴, Kristian Cibulskis⁵, Andrey Sivachenko³, Petar Stojanov⁴, Aaron McKenna⁴, Eric S. Lander^{6,7}, Stacey Gabriel⁸, Gad Getz⁴; **Genome Sequencing Center Washington University in St Louis** Li Ding^{9,10}, Robert S. Fulton⁹, Daniel C. Koboldt⁹, Todd Wylie⁹, Jason Walker⁹, David J. Dooling^{9,10}, Lucinda Fulton⁹, Kim D. Delehaunt⁹, Catrina C. Fronick⁹, Ryan Demeter⁹, Elaine R. Mardis^{9–11}, Richard K. Wilson^{9–11}; **Genome Characterization Center BC Cancer Agency** Andy Chu¹², Hye-Jung E. Chun¹², Andrew J. Mungall¹², Erin Pleasance¹², A. Gordon Robertson¹², Dominik Stoll¹², Miruna Balasundaram¹², Inanc Birol¹², Yaron S. N. Butterfield¹², Eric Chuah¹², Robin J. N. Coope¹², Noreen Dhalla¹², Ranabir Guin¹², Carrie Hirst¹², Martin Hirst¹², Robert A. Holt¹², Darlene Lee¹², Haiyan L. Li¹², Michael Mayo¹², Richard A. Moore¹², Jacqueline E. Schein¹², Jared R. Slobodan¹², Angela Tam¹², Nina Thiessen¹², Richard Varhol¹², Thomas Zeng¹², Yongjun Zhao¹², Steven J. M. Jones¹², Marco A. Marra¹²; **Genome-Characterization Center Broad Institute** Adam J. Bass^{4,13}, Alex H. Ramos^{4,13}, Gordon Saksena⁴, Andrew D. Cherniack⁴, Stephen E. Schumacher^{4,13}, Barbara Tabak^{4,13}, Scott L. Carter^{4,13}, Nam H. Pho⁴, Huy Nguyen⁴, Robert C. Onofrio⁴, Andrew Crenshaw⁴, Kristin Ardlie⁴, Rameen Beroukhi^{4,13}, Wendy Winkler⁴, Gad Getz⁴, Matthew Meyerson^{4,13,14}; **Genome-Characterization Center Brigham and Women's Hospital and Harvard Medical School** Alexei Protopopov¹⁵, Junhua Zhang¹⁵, Angela Hadjipanayis^{16,17}, Eunjung Lee^{17,18}, Ruibin Xi¹⁸, Lixing Yang¹⁸, Xiaojia Ren¹⁵, Hailei Zhang¹⁵, Narayanan Sathiamoorthy¹⁹, Sachet Shukla¹⁵, Peng-Chieh Chen^{16,17}, Psalm Haseley^{17,18}, Yonghong Xiao¹⁵, Semin Lee¹⁸, Jonathan Seidman¹⁶, Lynda Chin^{4,15,20}, Peter J. Park^{17–19}, Raju Kucherlapati^{16,17}; **Genome-Characterization Center University of North Carolina, Chapel Hill** J. Todd Auman^{21,22}, Katherine A. Hoadley^{23–25}, Ying Du²⁵, Matthew D. Wilkerson²⁵, Yan Shi²⁵, Christina Liqiori²⁵, Shaowu Meng²⁵, Ling Li²⁵, Yidi J. Turman²⁵, Michael D. Topal^{24,25}, Donghui Tan²⁶, Scot Waring²⁵, Elizabeth Buda²⁵, Jesse Walsh²⁵, Corbin D. Jones²⁷, Piotr A. Mieczkowski²³, Darshan Singh²⁸, Junyuan Wu²⁵, Anisha Gulabani²⁵, Peter Dolina²⁵, Tom Bodenheimer²⁵, Alan P. Hoyle²⁵, Janae V. Simons²⁵, Matthew Soloway²⁵, Lisle E. Mose²⁴, Stuart R. Jefferys²⁴, Saianand Balu²⁵, Brian D. O'Connor²⁵,

Jan F. Prins²⁸, Derek Y. Chiang^{23,25}, D. Neil Hayes^{25,29}, Charles M. Perou^{23–25}; **Genome-Characterization Centers University of Southern California and Johns Hopkins University** Toshinori Hinoue³⁰, Daniel J. Weisenberger³⁰, Dennis T. Maglinte³⁰, Fei Pan³⁰, Benjamin P. Berman³⁰, David J. Van Den Berg³⁰, Hui Shen³⁰, Timothy Triche Jr³⁰, Stephen B. Baylin³¹, Peter W. Laird³⁰; **Genome Data Analysis Center Broad Institute** Gad Getz⁴, Michael Noble⁴, Doug Voet⁴, Gordon Saksena⁴, Nils Gehlenborg^{4,18}, Daniel DiCara⁴, Junhua Zhang^{4,15}, Hailei Zhang^{4,15}, Chang-Jiun Wu^{4,15}, Spring Yingchun Liu^{4,15}, Sachet Shukla^{4,15}, Michael S. Lawrence⁴, Lihua Zhou⁴, Andrey Sivachenko⁴, Pei Lin⁴, Petar Stojanov⁴, Rui Jing⁴, Richard W. Park¹⁸, Marc-Danie Nazaire⁴, Jim Robinson⁴, Helga Thorvaldsdottir⁴, Jill Mesirov⁴, Peter J. Park^{17–19}, Lynda Chin^{4,15,20}; **Genome Data Analysis Center Institute for Systems Biology** Vesteinn Thorsson³², Sheila M. Reynolds³², Brady Bernard³², Richard Kreisberg³², Jake Lin³², Lisa Iype³², Ryan Bressler³², Timo Erkkila³², Madhumati Gundapuneni³², Yuxin Liu³³, Adam Norberg³², Tom Robinson³², Da Yang³³, Wei Zhang³³, Ilya Shmulevich³²; **Genome Data Analysis Center Memorial Sloan-Kettering Cancer Center** Jorma J. de Ronde^{34,35}, Nikolaus Schultz³⁴, Ethan Cerami³⁴, Giovanni Ciriello³⁴, Arthur P. Goldberg³⁴, Benjamin Gross³⁴, Anders Jacobsen³⁴, Jianjiong Gao³⁴, Bogumil Kaczkowski³⁴, Rileen Sinha³⁴, B. Arman Aksoy³⁴, Yevgeniy Antipin³⁴, Boris Reva³⁴, Ronglai Shen³⁶, Barry S. Taylor³⁴, Timothy A. Chan³⁷, Marc Ladanyi³⁸, Chris Sander³⁴; **Genome Data Analysis Center Institute of Texas MD Anderson Cancer Center** Rehan Akbani³⁹, Nianxiang Zhang³⁹, Bradley M. Broom³⁹, Tod Casasent³⁹, Anna Unruh³⁹, Chris Wakefield³⁹, Stanley R. Hamilton³³, R. Craig Cason³³, Keith A. Baggerly³⁹, John N. Weinstein^{39,40}; **Genome Data Analysis Centers, University of California, Santa Cruz and the Buck Institute** David Haussler^{41,42}, Christopher C. Benz⁴³, Joshua M. Stuart⁴¹, Stephen C. Benz⁴¹, J. Zachary Sanborn⁴¹, Charles J. Vaske⁴¹, Jingchun Zhu⁴¹, Christopher Szeto⁴¹, Gary K. Scott⁴³, Christina Yau⁴³, Sam Ng⁴¹, Ted Goldstein⁴¹, Kyle Ellrott⁴¹, Eric Collisson⁴⁴, Aaron E. Cozen⁴¹, Daniel Zerbino⁴¹, Christopher Wilks⁴¹, Brian Craft⁴¹, Paul Spellman⁴⁵; **Biospecimen Core Resource International Genomics Consortium** Robert Penny⁴⁶, Troy Shelton⁴⁶, Martha Hatfield⁴⁶, Scott Morris⁴⁶, Peggy Yena⁴⁶, Candace Shelton⁴⁶, Mark Sherman⁴⁶, Joseph Paulauskis⁴⁶; **Nationwide Children's Hospital Biospecimen Core Resource** Julie M. Gastier-Foster^{47–49}, Jay Bowen⁴⁷, Nilsa C. Ramirez^{47,48}, Aaron Black⁴⁷, Robert Pyatt^{47,48}, Lisa Wise⁴⁷, Peter White^{47,49}; **Tissue source sites and disease working group** Monica Bertagnolli⁵⁰, Jen Brown⁵¹, Timothy A. Chan⁵², Gerald C. Chu⁵³, Christine Czerwinski⁵¹, Fred Denstman⁵⁴, Rajiv Dhir⁵⁵, Arnulf Dörner⁵⁶, Charles S. Fuchs^{57,58}, Jose G. Guillem⁵⁹, Mary Iacocca⁵¹, Hartmut Juhl⁶⁰, Andrew Kaufman⁵², Bernard Kohl III⁶¹, Xuan Van Le⁶¹, Maria C. Mariano⁶², Elizabeth N. Medina⁶², Michael Meyers⁶³, Garrett M. Nash⁵⁹, Phillip B. Paty⁵⁹, Nicholas Petrelli⁵⁴, Brenda Rabeno⁵¹, William G. Richards⁵⁴, David Solit⁶⁶, Pat Swanson⁵¹, Larissa Temple⁵², Joel E. Tepper⁶⁵, Richard Thorp⁶¹, Efsevia Vakiani⁶², Martin R. Weiser⁵⁹, Joseph E. Willis⁶⁷, Gary Witkin⁵¹, Zhaosheng Zeng⁵⁹, Michael J. Zinner⁶³, Carsten Zornig⁶⁸; **Data-Coordination Center** Mark A. Jensen⁶⁹, Robert Sfeir⁶⁹, Ari B. Kahn⁶⁹, Anna L. Chu⁶⁹, Prachi Kothiyal⁶⁹, Zhining Wang⁶⁹, Eric E. Snyder⁶⁹, Joan Ponting⁶⁹, Todd D. Pihl⁶⁹, Brenda Ayala⁶⁹, Mark Backus⁶⁹, Jessica Walton⁶⁹, Jon Whitmore⁶⁹, Julien Baboud⁶⁹, Dominique L. Berton⁶⁹, Matthew C. Nicholls⁶⁹, Deepak Srinivasan⁶⁹, Rohini Raman⁶⁹, Stanley Girshik⁶⁹, Peter A. Kigonya⁶⁹, Shelley Alonso⁶⁹, Rashmi N. Sanbhadri⁶⁹, Sean P. Barletta⁶⁹, John M. Greene⁶⁹, David A. Pot⁶⁹; **Project Team National Cancer Institute** Kenna R. Mills Shaw⁷⁰, Laura A. L. Dillon⁷⁰, Ken Buetow⁷¹, Tanja Davidsen⁷¹, John A. Demchok⁷⁰, Greg Eley⁷², Martin Ferguson⁷³, Peter Fielding⁷⁰, Carl Schaefer⁷¹, Margi Sheth⁷⁰ and Liming Yang⁷⁰; **Project Team National Human Genome Research Institute** Mark S. Guyer⁷⁴, Bradley A. Ozenberger⁷⁴, Jacqueline D. Palchik⁷⁴, Jane Peterson⁷⁴, Heidi J. Sofia⁷⁴ & Elizabeth Thomson⁷⁴.

¹Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. ²Department of Biology and Biochemistry, University of Houston, Houston, Texas 77204, USA. ³Dan L. Duncan Cancer Center, Human Genome Sequencing Center, Baylor College of Medicine, Houston, Texas 77030, USA. ⁴The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. ⁵Medical Sequencing Analysis and Informatics, The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. ⁶Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. ⁷Department of Systems Biology, Harvard University, Boston, Massachusetts 02115, USA. ⁸Genetic Analysis Platform, The Eli and Edythe L. Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts 02142, USA. ⁹The Genome Institute, Washington University School of Medicine, St Louis, Missouri 63108 USA. ¹⁰Department of Genetics, Washington University School of Medicine, St Louis, Missouri 63108, USA. ¹¹Site-man Cancer Center, Washington University School of Medicine, St Louis, Missouri 63108, USA. ¹²Canada's Michael Smith Genome Sciences Centre, BC Cancer Agency, Vancouver, British Columbia V5Z 1L3, Canada. ¹³Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ¹⁴Department of Pathology, Harvard Medical School, Boston, Massachusetts 02115, USA. ¹⁵Belfer Institute for Applied Cancer Science, Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, USA. ¹⁶Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA. ¹⁷Division of Genetics, Brigham and Women's Hospital, Boston, Massachusetts 02115, USA. ¹⁸The Center for Biomedical Informatics, Harvard Medical School, Boston, Massachusetts 02115, USA. ¹⁹Informatics Program, Children's Hospital, Boston, Massachusetts 02115, USA. ²⁰Department of Dermatology, Harvard Medical School, Boston, Massachusetts 02115, USA. ²¹Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ²²Institute for Pharmacogenetics and Individualized Therapy, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ²³Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ²⁴Department of Pathology and Laboratory Medicine, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599,

USA. ²⁵Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ²⁶Carolina Center for Genome Sciences, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ²⁷Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ²⁸Department of Computer Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ²⁹Department of Internal Medicine, Division of Medical Oncology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA. ³⁰University of Southern California Epigenome Center, University of Southern California, Los Angeles, California 90089 USA. ³¹Cancer Biology Division, The Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, Maryland 21231, USA. ³²Institute for Systems Biology, Seattle, Washington 98109, USA. ³³Division of Pathology and Laboratory Medicine, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. ³⁴Computational Biology Center, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. ³⁵Divisions of Experimental Therapy, Molecular Biology, Surgical Oncology, The Netherlands Cancer Institute, 1066 CX Amsterdam, The Netherlands. ³⁶Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. ³⁷Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. ³⁸Department of Pathology, Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. ³⁹Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. ⁴⁰Department of Systems Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA. ⁴¹Department of Biomolecular Engineering and Center for Biomolecular Science and Engineering, University of California Santa Cruz, Santa Cruz, California 95064, USA. ⁴²Howard Hughes Medical Institute, University of California Santa Cruz, Santa Cruz, California 95064, USA. ⁴³Buck Institute for Age Research, Novato, California 94945, USA. ⁴⁴Division of Hematology/Oncology, University of California San Francisco, San Francisco, California 94143, USA. ⁴⁵Oregon Health and Science University, Department of Molecular and Medical Genetics, Portland, Oregon 97239, USA. ⁴⁶International Genomics Consortium, Phoenix, Arizona 85004, USA. ⁴⁷Nationwide Children's Hospital Biospecimen Core Resource, The Research Institute at Nationwide Children's Hospital, Columbus, Ohio 43205, USA. ⁴⁸The Ohio State University College of Medicine, Department of Pathology, Columbus, Ohio 43205, USA. ⁴⁹The Ohio State University College of Medicine, Department of Pediatrics, Columbus, Ohio 43205, USA. ⁵⁰Department of Surgery, Brigham and Women's Hospital, Harvard Medical School, Brookline, Massachusetts 02115, USA. ⁵¹Department of Pathology, Christiana Care Health Services, Newark, Delaware 19718, USA. ⁵²Human Oncology and Pathogenesis Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. ⁵³Department of Pathology, Brigham and Women's Hospital, Harvard Medical School, Brookline, Massachusetts 02115, USA. ⁵⁴Department of Surgery, Helen F. Graham Cancer Center at Christiana Care, Newark, Delaware 19718, USA. ⁵⁵Department of Pathology, University of Pittsburgh, Pittsburgh, Pennsylvania 15213, USA. ⁵⁶Klinik für Chirurgie, Krankenhaus Alten Eichen, 22527 Hamburg, Germany. ⁵⁷Department of Medical Oncology, Dana-Farber Cancer Institute, Brookline, Massachusetts 02115, USA. ⁵⁸Department of Medicine, Brigham and Women's Hospital, Brookline, Massachusetts 02115, USA. ⁵⁹Department of Surgery, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. ⁶⁰Indivumed Inc., Kensington, Maryland 20895, USA. ⁶¹ILSbio, LLC, Chestertown, Maryland 21620, USA. ⁶²Department of Pathology, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. ⁶³Department of Surgery, Brigham and Women's Hospital, Brookline, Massachusetts 02115, USA. ⁶⁴Tissue and Blood Repository, Brigham and Women's Hospital, Brookline, Massachusetts 02115, USA. ⁶⁵Dept of Radiation Oncology, University of North Carolina School of Medicine, Chapel Hill, North Carolina 27599, USA. ⁶⁶Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA. ⁶⁷Department of Pathology, Case Medical Center, Cleveland, Ohio 44106, USA. ⁶⁸Chirurgische Klinik, Israelitisches Krankenhaus, 22297 Hamburg, Germany. ⁶⁹SRA International, Fairfax, Virginia 22033, USA. ⁷⁰The Cancer Genome Atlas Program Office, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892, USA. ⁷¹Center for Biomedical Informatics and Information Technology (CBIIIT), National Cancer Institute, National Institutes of Health, Rockville, Maryland 20852, USA. ⁷²Scimentis, LLC, Statham, Georgia 30666, USA. ⁷³MLF Consulting, Arlington, Massachusetts 02474, USA. ⁷⁴National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland 20892, USA.