doi:10.1038/nature11252

Supplement Contents

gene clusters. Contains the results of the Hypergeometric overlap test in which constituent pathways from the SuperPathway were overlapped with each of the 50 gene clusters. Overlap statistics show the -log(p-value) ("Significance" column), the number of genes in common between the constituent pathway and the cluster ("Overlap" column), the number of genes in the cluster ("Cluster Size" column), the number of genes in the constituent pathway ("Pathway Size" column), and the total number of genes in the SuperPathway ("Total Genes" column). Worksheet B: All of the genes listed as they appear in rows of the heatmap depicted in Figure 5A of the main text ("GENE" Column; HUGO symbol). Only genes within the SuperPathway were used for hierarchical clustering and are listed in cluster order. Pathway indicates to which constituent pathway, used to build the SuperPathway, the gene belongs. If a gene belonged to multiple constituent pathways the one most enriched in the cluster was selected. Worksheet C: All of the samples listed as they appear in the columns of the heatmap in Figure 5A.

- **Supplementary Table 10.** Gene Expression, SCNA, and MicroRNA signatures associated with CRC Tumor Aggression. Signatures are displayed that meet a statistical threshold for the combined p-value by the weighted Fisher’s method after multiple test corrections ($p<1\ 10^{-7}$,gene expression; $p<10^{-4}$ SCNA, $p < 0.001$ MicroRNAs), and are also significant in the subset of samples that are either MSS or MSI-L. Column “Agg.” is 1 for a signature that is elevated in aggressive tumors, and -1 for a signature that is lower. Remaining columns: p-values for individual tests, no multiple testing correction. (More extensive tables, including those for methylation, mutation and Paradigm signatures, are provided in Supplementary Table 9).
- **Supplementary Table 11.** Gene Expression, SCNA, and MicroRNA, mutation and Paradigm signatures associated with CRC Tumor Aggression.
- **Supplementary Table 12.** Epigenetically silenced genes in colorectal cancer

Supplementary Figures

- **Supplementary Figure 1.** DNA methylation-based subgroups in CRC and their molecular and clinical features. We performed unsupervised clustering to determine subgroups in colorectal tumors based on their promoter DNA methylation profiles (See Supplemental Methods section). Shown is heatmap representation of DNA methylation β-values of 1,403 most variable probes (standard deviation >0.20) across 236 tumor samples, with dark blue indicating low DNA methylation and yellow indicating high DNA methylation. The RPMM-based cluster assignments are indicated above the heatmap: light sky blue, CIMP-H (n=36); orchid, CIMP-L (n=53); gold, cluster 3 (n=77) and light green, cluster 4 (n=70). Selected molecular and clinical features of each tumor sample are also shown as color bars above the heatmap, as indicated in the legends to the right of the heatmap.
- **Supplementary Figure 2.** Hierarchical clustering of microarray gene expression data in 220 colorectal tumors reveals three main clusters. The tumors can be classified into three groups represented at the bottom of the Figure.
- **Supplementary Figure 3.** Consensus heatmap for three sample groups identified from miRNA-seq abundance profiles for 255 colorectal tumor samples. NMF consensus clustering was applied to a normalized abundance matrix for the 25% most variant mature or star strands (221 MIMATs). The legend colors and dendrogram reflect per-sample cluster membership over 1000 iterations. Tracks under the heatmap show NMF clusters, colon vs. rectum sample types, then sample classifications from DNA methylation clusters: CIMP-H, CIMP-L, cluster 3 and cluster 4.

- **Supplementary Figure 4.** Somatic copy number variation deduced by single nucleotide polymorphism (SNP) arrays in 257 colon and rectal samples. **A.** Focal Peaks (the number of genes in each peak is shown in parentheses). **B.** Broad alterations per chromosome arm.
- **Supplementary Figure 5.** Supplementary Figure 5. Molecular basis for chromosomal translocations. For two of the recurrent translocations, the putative regions of breakpoints as deduced from the location of non-concordant reads (shown by arrows for *NAV2*-*TCF7L1*) were amplified and subjected to sequencing by capillary eletrophoresis. One such sequence from *NAV2*-*TCF7L1* fusions and two from translocations involving *TTC28* are shown. The numbers represent genomic coordinates. **A.** Translocations involving *NAV2* and *TCF7L1*. Three cases of translocations involving these two genes were found. In all cases the translocation breakpoints are in intron 3 of *NAV2* and intron 3 of *TCF7L1* (top). The locations of discordant read pairs supporting the translocation in each of the three tumors are shown. The translocation junction sequence derived from sequencing the amplified product from TCGA-AA-A00U is shown. Nucleotides in black correspond to those from *NAV2* and those in red are from *TCF7L1*. **B.** Sequence derived from amplified products from two translocations involving *TTC28*. The translocation involving *TTC17* and *TTC28* is a simple translocation while the translocation involving *SPATA16* and *TTC28* is complex and involves three chromosomes.
- **Supplementary Figure 6.** Genomic alteration patterns in select pathways in the non-hypermutated tumors. Each column of an OncoPrint represents an individual case, each row represents a gene (see Supplementary Table 1 for a full table of all events). Only cases with a pathway alteration are shown, and only the 165 non-hypermutated samples were included in this analysis.
- **Supplementary Figure 7.** FBXW7 CRC mutations in the context of protein structure. Missense mutations in FBXW7 (green) are located almost entirely in the core of the target protein binding beta-propeller. These mutations directly disrupt the ability of FBXW7 to bind to and catalyze ubiquitination of target proteins. The most frequent FBXW7 mutation is circled at position R465. Similarly, nonsense and frameshift mutations (red) affect the C-terminal protein binding beta-propeller exclusively, thereby disrupting FBXW7-catalyzed ubiquitination.
- **Supplementary Figure 8.** Altered Patterns of Expression, Copy Number, Methylation and Mutations Associated with Tumor Aggressiveness.
  A. Expression of Gene *APOL6* is Significantly Lower in Aggressive Tumors. *APOL6* gene expression (composite *p*-value $p<10^{-16}$) is shown for individual clinical data values. The *APOL6* colorbar from Figure 5B is reproduced, and examples are given of individual clinical associations contributing to that colorbar. Each subpanel shows how *APOL6* expression levels segregate by clinical variable, along with the (unadjusted) individual *p*-values and color. The other individual *p*-values are $p=1.5\bullet10^{-4}$ for Positive Lymph Nodes and $p=0.012$ for Vascular Invasion. *APOL6*, a member of the apolipoprotein L family, induces mitochondria mediated apoptosis in colorectal cancer cells [PMID:15671246].
  B. Clinical Associations for Mutations in Key Colorectal Cancer Proteins. PIK3CA, FBXW7 and BRAF mutations are less prevalent than expected in aggressive tumors, whereas TP53 and APC mutations show the opposite behavior. Association can be specifically dependent on protein domain context. APC mutations occur in several defined regions in the protein, but only those found early in the protein, leading to loss of Armadillo repeats, show strong clinical association (e.g. Armadillo Repeat 1: $p=3\bullet10^{-7}$). In comparison, APC non-silent mutations as a whole lack association with tumor aggressiveness (combined $p=0.76$; all individual clinical comparisons have $p>0.1$). TP53 mutants also exhibit a shift in clinical associativity depending on the possible locations of

mutations within specific domains. Aggressiveness associations with BRAF and TP53 are largely dictated by sample segregation by hypermutation and MSI status. (Mutations leading to loss of TP53 binding domain are found almost exclusively in non-hypermutated samples). Threshold for inclusion in figure: *p*=0.05 and 10 mutations. **C.** Tumor Aggressiveness Markers in the chromosome 22 region q12.3-13.2. The region includes apolipoprotein L family members (APOL1-4,6), immune receptor genes *IL2RB* (composite *p*=$1.6\bullet10^{-7}$) and *CSF2RB* (*p*=$2.2\bullet10^{-4}$) and apolipoprotein B mRNA editing enzyme family member *APOBEC3* all of which show decreased expression in aggressive tumors. The region also includes mutations in the gene *EP300*, (*p*=$2\bullet10^{-4}$) coding for the transcriptional co-activator p300, which have a prior association with colorectal cancer [PMID:1473269].

- **Supplementary Figure 9.** Integrated large-scale analysis reveals the presence of multiple genomic regions that dictate tumor aggressiveness. **A.** All molecular signatures, including gene expression, probe-level methylation, somatic copy number alterations, microRNAs, and gene mutation frequencies were scored on the basis of their combined statistical association with the clinical measurements of histological type, metastasis, tumor stage, fraction of positive lymph nodes, and vascular and lymphatic invasion. Molecular signatures significantly associated with tumor aggression (combined p-value $p<10^{-3}$) are displayed as tiles according to data type (outer ring), and on a scale of tumor aggressiveness (inner ring). Inner ring: Molecular readouts that are elevated in more aggressive tumors are shown in red, and those trending oppositely are blue, with the color intensity indicating the strength of the association. A web-based tool that allows interactive exploration of clinically correlated regions is available at explorer.cancerregulome.org. **B.** Detailed view of chromosome 20. Certain chromosomal regions are enriched in clinically associated molecular features. Region 20q13.12 includes a local amplification (orange) and 11 genes (blue), all of which are expressed more highly in aggressive tumors. A number of methylation probes (green) are also statistically associated with tumor aggression, nearly all (8/10) with decreased levels in aggressive tumors.

Supplementary Data at http://tcga-data.nci.nih.gov/docs/publications/coadread_2012/

- **Supplementary Data File 1.** A ZIP archive file containing the relevant data to reproduce the PARADIGM pathway analysis. The archive contains the following five files:
    - **SuperPathway.txt**: Superimposed Pathway used by the PARADIGM analysis. All of the merged concepts and interactions pooled from NCI-PID, Reactome, and BioCarta databases. At the top of the file, declarations of all of the concepts (genes, complexes, families, processes) can be found. Beneath these declarations are all of the regulatory interactions including transcriptionally activating (-t>), transcriptionally inactivating (-t|), subunit to complex relations (-component>), post-transcriptionally activating (-a>), post-transcriptionally inactivating (-a|), activation of an abstract process (-ap>), inhibition of an abstract process (-ap|), and membership in a family relation (-member>).
    - **tcgaCOADREAD_Expression.vNormal.MANUSCRIPT.tab**: A PARADIGM-ready version of the expression data formatted as a tab-delimited file with the expression rank-ratios given as input to the PARADIGM algorithm.
    - **tcgaCOADREAD_CNV.vNormal.MANUSCRIPT.tab**: A PARADIGM-ready version of the copy number data. A tab-delimited file containing the copy number rank-ratios given as input to the PARADIGM algorithm.

  - **params.txt**: The set of parameters needed to run PARADIGM that determine the initial setting of the constraints between concept- and interaction-related constraints (probabilistic factors). These parameters were learned from previous rounds of learning on other cancer cohorts and reused for this analysis.
  - **config.txt**: Contain settings for how PARADIGM's inference engine was run for the CRC analysis. The file specifies that the belief propagation method for maximum likelihood inference should be used with a maximum of 10,000 iterations for convergence and that the datasets for gene expression and copy number to be used are the files listed above.
- **Supplementary Data File 2. A network of the pathway concepts found by PARADIGM to be significantly modulated across the colonic and rectal tumor samples.** The file **modulated.cys** contains the network as a Cytoscape session that has been tested on versions 2.6 or later. Nodes in the network correspond to concepts in the Superimposed Pathway and include genes (circles), complexes (hexagons), families (triangles), and cellular processes (boxes). Concepts are connected by regulatory interactions depicted as either activating (arrows) or inhibiting ("T"-bars) at the transcriptional level (solid lines), or post-transcriptional level (dashed lines). Subunit membership in complexes is depicted using undirected dashed lines. The network includes concepts with higher activation (red nodes) or inactivation (blue nodes) in tumors compared to normal. The size and opacity of the nodes are drawn as a function of the modulation score.

Supplementary Methods

Specimen Samples

Sample inclusion criteria. Biospecimens were collected from newly diagnosed patients with colon or rectum adenocarcinoma undergoing surgical resection and had received no prior treatment for their disease, including chemotherapy or radiotherapy. All cases were collected regardless of surgical stage or histologic grade. Cases were staged according to the American Joint Committee on Cancer (AJCC) staging system. Each frozen tumor specimen had a companion normal tissue specimen which could be blood/blood components, adjacent normal tissue taken from greater than 2cm from the tumor, or previously extracted germline DNA from blood. Each tumor specimen weighed at least 60 mg and was typically under 200 mg. Specimens were shipped overnight from two tissue source sites (Indivumed and Christiana Care) using a cryoport that maintained an average temperature of less then -180°C. Each tumor and adjacent normal tissue specimen were embedded in optimal cutting temperature (OCT) medium and histologic sections were obtained from top and bottom portions for review. Each H&E stained case was reviewed by a board-certified pathologist to confirm that the tumor specimen was histologically consistent with colon adenocarcinoma and the adjacent normal specimen contained no tumor cells. The sections were required to contain an average of 60% tumor cell nuclei with less than 20% necrosis for inclusion in the study per TCGA protocol requirements.

RNA and DNA were extracted from tumor specimens using a modification of the DNA/RNA AllPrep kit (Qiagen). The isolation methodology for each sample was noted in the Biospecimen XML uploaded to the DCC (http://tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp). A portion of the flow-through from the DNA column was processed according to the AllPrep RNA extraction instructions to produce RNA analytes >200 nt [designated 'Allprep RNA Extraction' in the biospecimen XML], while the other portion was either precipitated after TRIzol separation [designated 'Total RNA' in the XML] or purified using a *mir*Vana miRNA Isolation Kit (Ambion). This latter step generated RNA preparations that included RNA <200 nt [designated 'mirVana (Allprep DNA) RNA'] suitable for miRNA analysis. DNA was extracted from normal tissue using either the QiaAmp blood midi kit (Qiagen) or the QiaAmp tissue mini kit (Qiagen). Each specimen was quantified by measuring Abs_{260} with a UV spectrophotometer. DNA specimens were resolved by agarose gel electrophoresis to determine the range of fragment sizes. The AmpFISTR Identifiler (Applied Biosystems) or Sequenom SNP panel procedure was utilized to verify tumor DNA and germline DNA were derived from the same patient. One µg each of tumor and normal DNA was sent to Qiagen for REPLI-g whole genome amplification using a 100 µg reaction scale. Only those specimens yielding a minimum of 6.9 µg of tumor DNA, 2.15 µg of column-purified RNA, 3.15 µg TRIzol precipitated or *mir*Vana-purified RNA, and 4.9 µg of germline DNA were included in this study. Total and Column-purified RNA was analyzed via the RNA6000 assay (Agilent) for determination of an RNA Integrity Number (RIN), and only the cases with RIN >7.0 were included in this study. At the time of the data freeze, 345 colon/rectum adenocarcinoma cases were received by the BCRs and 56% passed quality control.

Microsatellite instability testing. Microsatellite instability (MSI) status of the adenocarcinomas was evaluated in the clinical Molecular Diagnostics Laboratory of the Division of Pathology and Laboratory Medicine at The University of Texas M. D. Anderson Cancer Center and in the Biospecimen Core Resource at Nationwide Children's Hospital. At M.D. Anderson, a panel of four mononucleotide repeat sequences (polyadenine tracts BAT25, BAT26, BAT40, and transforming growth factor receptor type II) and three dinucleotide repeat sequences (CA repeats in D2S123, D5S346, & D17S250) was used on the basis of the recommendations from the National Cancer Institute Workshop on MSI in 2002[1]. At Nationwide Children's Hospital, the MSI Analysis System, Version 1.2 (Promega, Madison, WI) was used according to manufacturer's instructions. This system uses five mononucleotide repeat markers (BAT-25, BAT-26, NR-21, NR-24, and MONO-27) for MSI interpretation and two pentanucleotide repeat markers (Penta C and Penta) to confirm sample identity. Electrophoretic mobility in these microsatellites from 202 tumors and matched non-neoplastic tissue or mononuclear blood cells was compared after multiplex fluorescent-labeled PCR and capillary electrophoresis to identify variation in the number of repeats. Equivocal or failed markers were re-evaluated by singleplex PCR or through re-analysis of the entire MSI panel. Altered size of no marker in tumor DNA resulted in classification of the tumor as microsatellite-stable (MSS), one or two altered markers (<30%) as low levels of MSI (MSI-L), three altered markers (43%) as equivocal, and five to seven altered markers (>70%) as high levels of MSI (MSI-H). No case had alteration of four markers. In all MSI-L tumors and the one equivocal case, an additional six dinucleotide repeats on chromosome 18q (D18S69, 64, 1147, 55, 61 & 58 from centromere to telomere) were analyzed, and all tumors were MSI-L.

Data Coordination Center (DCC)

Data flow and organization. The TCGA data and analysis network involves leaders and teams working at multiple institutions and organizations in defined and coordinated roles, described briefly in the following overview. The Biospecimen Core Resources (BCRs) receive tissue samples and clinical metadata from Tissue Source Sites which are independently contracted by TCGA. The BCRs extract biospecimen analytes (DNA and RNA) from tissue samples, and ship plates containing aliquots of these analytes to the TCGA Genomic Sequencing Centers (GSCs) and Genomic Characterization Centers (GCC). BCRs assign each aliquot a unique and persistent biospecimen identifier, from which the parent analyte, sample, and case may be traced. These identifiers accompany aliquot shipments to the GSCs and GCCs. Data from any aliquot produced and submitted by the GSCs and GCCs are permanently associated with the aliquot identifier. The BCRs transfer biospecimen and clinical metadata along with the identifiers to the TCGA Data Coordinating Center (DCC), which maintains the identifier-metadata associations in a custom relational database. Genomic Data Analysis Centers (GDACs) perform integrative analyses across datatypes. GDAC results and applications are currently managed directly by these centers.

GSCs, as well as GCCs producing RNA-Seq data, submitted sequence data, aligned to the hg18 reference genome, to NCBI's Database of Genotype and Phenotype (dbGaP); sequence data were maintained in NCBI's Sequence Read Archive (SRA). All other data, including instrument data for array-based platforms, and normalized and/or interpreted data (e.g., mutation calls or relative expression values), were submitted to the DCC. In general, GSC data available at the DCC include called and annotated somatic mutations in custom Mutation Annotation Format (MAF) files, as well as enhanced Variant Calling Format (VCF) files. GCC data at the DCC include primary instrument data (level 1), normalized or otherwise initially processed data (level 2), interpreted or segmented data (level 3), and summary, region of interest, or cross-datatype integrated analyses (level 4). DCC-managed data incorporates gene expression calls (for array-based, RNA-Seq, and miRNA-Seq platforms), array-based SNP calls, copy number variation, loss of heterozygosity (LOH), and DNA methylation. Array-based data archives are accompanied by Investigation Description Format (IDF) and Sample-Data Relationship Format (SDRF) files. These files comply with the MAGE-TAB format standard. A complete list of TCGA instrument platforms and data types is given in Supplementary Methods Table 1.

TCGA data at the DCC is organized by platform, data type and data level. Each platform may produce several types of data. For example, SNP-based platforms yield three data types: SNP, Copy Number Results, and Loss of Heterozygosity (LOH). The TCGA concept of data level segregates raw data from derived data, and derived data from higher-level analysis or interpreted results. Each center and platform may have a slightly different concept of data level depending on their data types, platforms, and the algorithms used for analysis; therefore, the centers themselves make the assignment of data level with assistance from the DCC as necessary. General descriptions of TCGA data levels are provided in Supplementary Methods Tables 2 and 3.

Primary clinical and biospecimen data are submitted by the BCRs to the DCC in XML format. This XML is validated against public XML Schema documents that are deployed in directories at http://tcga-data.nci.nih.gov/docs/xsd/BCR/. The XML documents themselves are deployed by the DCC to the controlled access tier (see the section **Data Access**). The exact URLs of their associated schemas can be found in the header of the XML files themselves. For convenience, the DCC also provides flat, text-only tabular digests of clinical and biospecimen data elements that mirror the XML data, in which columns represent data elements and rows represent study cases. Tabular versions of the full clinical information are made available on the controlled access tier, while tabular versions of clinical information with PII removed are made available in the open access tier. Clinical and biospecimen terms referenced in the XML and tabular data are registered with the Cancer Data Standards Registry and Repository (caDSR) as clinical data elements. Clinical data element definitions and caDSR public identifiers can be found in the public TCGA Data Dictionary (http://tcga-data.nci.nih.gov/docs/dictionary/TCGA_BCR_DataDictionary.xml).

A more complete review of DCC data organization can be found on the TCGA Wiki, at the TCGA Data Primer (https://wiki.nci.nih.gov/x/j5dXAg) and links therein.

Data access. The DCC hosts a portal to TCGA data at http://tcga-data.nci.nih.gov. Applications for searching and downloading the data are available at this site. Extensive documentation on TCGA data organization and DCC applications is also maintained on the TCGA Wiki at https://wiki.nci.nih.gov/display/TCGA/TCGA+Wiki+Home. Questions about TCGA data and access may always be directed to TCGA-DCC-BINF-L@list.nih.gov.

The permanent set of data analyzed in the present paper has been collated and made available for direct download on the DCC portal.

Controlled data access policies and procedures. TCGA produces large volumes of genomic information derived from human tumor specimens collected from patient populations, and grants access to significant amounts of clinical information associated with these specimens. The aggregated data generated is unique to each enrolled case and, despite the lack of any direct identifying information within the data, there is a risk of individual re-identification by bioinformatic methods and/or third-party databases. Because patient privacy protection is paramount to NIH and TCGA, human subjects protection and data access policies are implemented to minimize the risk that the privacy of the donors and the confidentiality of their data will be compromised. As part of this effort, data generated from TCGA are available in two tiers.

The *open access* tier is publically accessible, and contains data that are considered by TCGA to present a low risk of re-identification of individual participants. The open access data tier does not require user certification for data access. The *controlled access* tier contains data, including all raw sequence data, that is unique to the individual participant or for which there is a high risk of individual re-identification as determined by TCGA. The controlled access tier encompasses all data classified by TCGA as Personally Identifiable Information (PII). This tier requires user certification for data access. For more information on these tiers and how to gain access to the controlled access tier, see http://tcga-data.nci.nih.gov/tcga/tcgaAccessTiers.jsp.

To administer these tiers, the DCC maintains two server branches. The open access branch (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/) houses open access data and is accessible without authentication. The controlled access branch (https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/tcga4yeo/) houses controlled data and is accessible by password authentication. Investigators who have applied for and received a Data Use Certificate (see http://tcga-data.nci.nih.gov/tcga/tcgaAccessTiers.jsp) can obtain a username and password. Sequence data at dbGaP is also password-accessible to investigators possessing a Data User Certificate.

The DCC also post-processes clinical and biospecimen data after submission and before deployment to remove certain PII. Post-processing consists of replacing any absolute dates (those dates specified by day, month, and year) with interval dates, specified by days since the date of initial pathologic diagnosis; negative values in the data indicate events prior to diagnosis date. This replacement occurs in both the XML and tabular formats, and affects both access tiers. The date of initial pathologic diagnosis is completely removed from deployed data; however, the year of diagnosis (the “index year”) is preserved. In addition, age-related fields of enrolled cases whose reported age is over 90 years are set to 90 years. This modification is indicated in the XML with an attribute “floored” set equal to “true”.

Data Sets

TCGA used single nucleotide polymorphism (SNP) arrays (Affymetrix) and low pass (3-5X coverage) whole genome sequencing (Illumina HiSeq 2000) to detect chromosome and sub-chromosomal copy number changes and translocations, microarray (Agilent) and RNA-Seq (Illumina) for mRNA expression profiling, Illumina Infinium HumanMethylation27 arrays to profile DNA methylation at gene promoters, miRNA quantification via Illumina sequencing and whole exome sequencing using both the Illumina and Solid platforms to detect coding mutations. Details about each of these platforms are presented below. Each of the platforms utilized a slightly different number of samples. Details of the sample IDs and the platforms used to study each of the samples are shown in Supplementary Table 1.

DNA Sequencing

Library construction: Illumina HiSeq. After QC, high molecular weight double strand genomic DNA samples are constructed into Illumina PairEnd precapture libraries according to the manufacturer’s protocol (Illumina Inc.) with modification. Briefly, 1ug genomic DNA in 100ul volume was sheared into fragments of approximately 300 base pairs in Covaris plate with E210 system (Covaris, Inc. Woburn, MA). The setting was 10% Duty cycle, Intensity of 4, 200 Cycles per Burst, for 120 seconds. Fragment size was checked using a 2.2 % Flash Gel DNA Cassette

(Lonza, Cat. No.57023). The Fragmented DNA was End-Repaired in 90ul total reaction volume containing sheared DNA, 9ul 10X buffer, 5ul END Repair Enzyme Mix and H2O (NEBNext End-Repair Module; Cat. No. E6050L) and then incubated at 20°C for 30 minutes. A-tailing was performed in a total reaction volume of 60ul containing End-Repaired DNA, 6ul 10X buffer, 3ul Klenow Fragment (NEBNext dA-Tailing Module; Cat. No. E6053L) and H_2O followed by incubation at 37°C for 30 minutes. Illumina multiplex adapter ligation (NEBNext Quick Ligation Module Cat. No. E6056L) was performed in a total reaction volume of 90ul containing 18ul 5X buffer, 5ul ligase, 0.5ul 100uM adaptor and H2O at room temperature for 30 minutes. After Ligation, PCR with Illumina PE 1.0 and modified barcode primers (manuscript in preparation) was performed in 170µl reactions containing 85 2x Phusion High-Fidelity PCR master mix, adaptor ligated DNA, 1.75ul of 50uM each primer and H2O. The standard thermocycling for PCR was 5' at 95°C for the initial denaturation followed by 6-10 cycles of 15 s at 95°C, 15 s at 60°C and 30 s at 72°C and a final extension for 5 min. at 72°C. Agencourt® XP® Beads (Beckman Coulter Genomics, Inc.; Cat. No. A63882) was used to purify DNA after each enzymatic reaction. After Beads purification, PCR product quantification and size distribution was determined using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517).

DNA sequencing: Illumina HiSeq. Sequencing was performed in paired-end mode with Illumina HiSeq 2000. Illumina sequencing libraries were amplified by "bridge-amplification" process using Illumina HiSeq pair read cluster generation kits (TruSeq PE Cluster Kit v2.5, Illumina) according to the manufacturer's recommended protocol. Briefly, these libraries were denatured with sodium hydroxide and diluted to 3-4 pM in hybridization buffer for loading onto a single lane of a flow cell in order to achieve 600-700k clusters/mm. All lanes were spiked with 1% phiX control library. Cluster formation, primer hybridization were performed on the flow cell with illumina's cBot cluster generation system.

Sequencing reactions were extended for 202 cycles of SBS using TruSeq SBS Kit on an Illumina's Hiseq 2000 sequencing machine according to the manufacturer's instructions. The Illumina Sequence Control Software (SCS) control the reagent delivery and collect raw images. Real Time Analysis (RTA) software was used to process the image analysis and base calling. On average, about 80-100 million successful reads, consisting of 2 X100 bp, were generated on each lane of a flow cell.

Library construction: SOLiD 4. Whole genome amplified (WGA) DNA samples (5ug) were constructed into SOLiD precature libraries according to a modified version of the manufacturer's protocol (Applied Biosystems, Inc.). Briefly, The genomic DNA was sheared into fragments of approximately 120 base pairs with the Covaris S2 or E210 system as per manufacturer instruction(Covaris, Inc. Woburn, MA). Fragments were processed through DNA End-Repair (NEBNext End-Repair Module; Cat. No. E6050L) and A-tailing (NEBNext dA-Tailing Module; Cat. No. E6053L). The resulting fragments were ligated with BCM-HGSC-designed Truncated-TA (TrTA) P1 and TA-P2 adapters with the NEB Quick Ligation Kit (Cat. No. M2200L). Solid Phase Reversible Immobilization (SPRI) bead cleanup (Beckman Coulter Genomics, Inc.; Cat. No. A29152) was used to purify the adapted fragments, after which nick translation and Ligation-Mediated PCR (LM-PCR) was performed using Platinum PCR Supermix HIFi (Invitrogen; Cat. No.12532-016) and 6 cycles of amplification. After Beads purification, PCR products' quantification and their size distribution were analyzed using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517). Primer sequences and a complete library construction protocol are available on the Baylor Human Genome

Website (http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf).

Exome capture and sequencing: SOLiD. The precapture libraries (2 ug) were hybridized in solution to NimbleGen CCDS Solution Probes which targets ~36 Mbs of sequence from ~17K genes, according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos TrTA-A and SOLiD-B replaced oligos PE-HE1 and PE-HE2 and post-capture LM-PCR was performed using 12 cycles. Capture libraries were quantified using PicoGreen (Cat. No. P7589) and their size distribution analyzed using the Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517). The efficiency of the capture was evaluated by performing a qPCR-based quality check on the built-in controls (qPCR SYBR Green assays, Applied Biosystems). Four standardized oligo sets, RUNX2, PRKG1, SMG1, and NLK, were employed as internal quality controls. The enrichment of the capture libraries was estimated to range from 7 to 9 fold over the background. The captured libraries were further processed for SOLiD sequencing. Primer sequences and a complete capture protocol are available on the Baylor Human Genome Website (http://www.hgsc.bcm.tmc.edu/documents/Preparation_of_SOLiD_Capture_Libraries.pdf)

Exome capture and sequencing: Illumina. Precapture libraries (1 ug) were hybridized in solution to NimbleGen SeqCap EZ Exome 2.0 Solution Probes targeting ~44Mbs of sequence from ~30K genes, or VCRome 2.1 (HGSC design, NimbleGen) targeting 43 Mb of sequence from ~30K genes, according to the manufacturer's protocol with minor revisions. Specifically, hybridization enhancing oligos IHE1, IHE2 and IHE3 (manuscript in preparation) replaced oligos HE1.1 and HE2.1 and post-capture LM-PCR was performed using 14 cycles. Capture libraries were quantified using Caliper GX 1K/12K/High Sensitivity Assay Labchip (Hopkinton, MA, Cat. No. 760517). The efficiency of the capture was evaluated by performing a qPCR-based quality check on the built-in controls (qPCR SYBR Green assays, Applied Biosystems). Four standardized oligo sets, RUNX2, PRKG1, SMG1, and NLK, were employed as internal quality controls. The enrichment of the capture libraries was estimated to range from 7 to 9 fold over background.

DNA Sequencing: SOLiD. Each captured library was hybridized to microbeads using Applied BioSystems' SOLiD platform-specific adapters) and submitted to an emulsion PCR to amplify the DNA fragments onto the beads (SOLiD ePCR Kit V2, Applied Biosystems). After amplification, the beads were recovered from the oil phase and the beads carrying amplified bound DNA were enriched (SOLiD Buffer and Bead Enrichment Kits, Applied Biosystems). The beads carrying amplified bound DNA were then modified to covalently adhere to a SOLiD coated slide (SOLiD Bead Deposition and Slide Kits, Applied Biosystems). The slides were loaded on the SOLiD v3 sequencing platform (SOLiD 3 Instrument Buffer Kit, Applied Biosystems) and sequenced over 8 days (SOLiD Fragment Library Sequencing Kit – MM50, Applied Biosystems).

Mapping Reads. *SOLiD.* Base and quality calling for SOLiD data was performed on-instrument using standard vendor software and settings. Upon completion of a run, read and quality data was copied into our data-center where individual sequence events are split into 10M read bundles and mapped in parallel using BFAST (version 0.6.4). After read bundles are mapped their results are merged back into a single sequence-event-level BAM where read group tags are added. Where necessary, sample-level BAMs are generated by merging using Picard (version 1.7), and duplicate reads are marked at the library level using SAMtools (version

1.7). Variant calling is done using custom filters applied to pileups made at the sample level, also using SAMtools.
Illumina. The output of a Illumina HiSeq sequencer are binary bcl files that are processed using the software (BCLConvertor 1.7.1). All reads from the prepared libraries that passed the illumina Chastity filter were formatted into fastq files. The fastq files are aligned to the genome using BWA (bwa-0.5.9rcl) against human genome build #18. The BWA reference library is constructed by the Broad Institute and then distributed to TCGA sequencing centers to ensure compatibility of aligned sequences among centers. Default parameters are used for alignment except for a 40 bp seed sequence, 2 mismatches in the seed, and a total of 3 mismatches allowed.

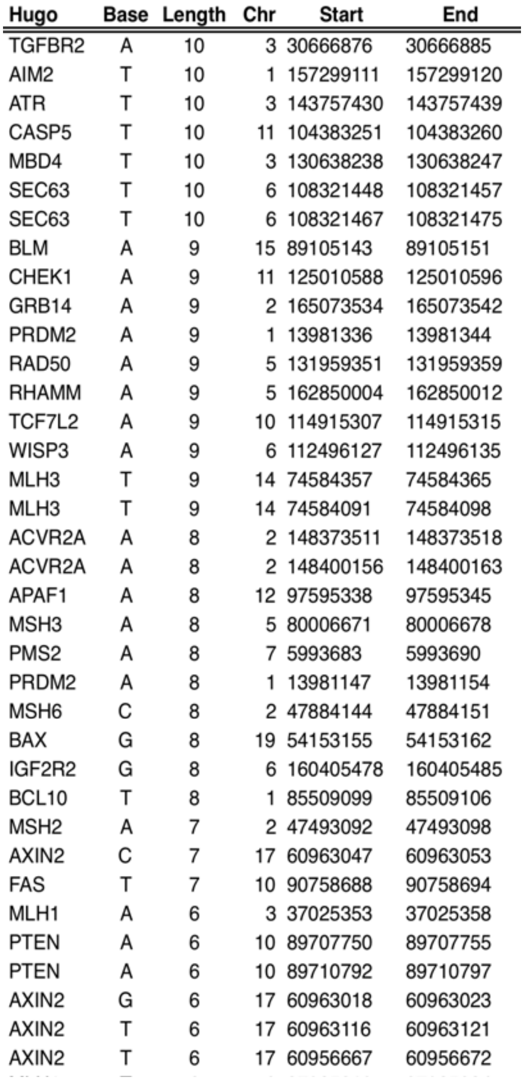
Mutation Detection. BAM files generated from alignment of Illumina sequencing reads were preprocessed using GATK. Mutations in Illumina data were discovered by the MuTect algorithm[2] (see also http://www.broadinstitute.org/cancer/cga/MuTect). Mutations in BAM files generated from SOLiD reads were detected as follows: SamTools Pileup was run to list all variants found in multiple reads at a single locus. The variants were further filtered to remove all those observed fewer than 5 times or were present in less than 0.10 of the reads. At least one variant had to be Q30 or better, and the variant had to lie in the central portion of the read, 15% from the 5' end of the read and 20% from the 3' end. In addition reads harboring the variant must have been observed in both forward and reverse orientations. Finally, the variant base was not observed in the normal tissue. Insertion or deletion variants ("indels") were discovered by similar processing except indels must have been observed in 0.25 of the reads (see below for detection of frameshift indels at microsatellite sites).

Validation of Mutations. Mutations were validated by running a second sequencing reaction on captured or PCR amplified DNA from the mutated sample and its matched normal. Since whole genome amplified DNA had been used in the mutation discovery phase, template DNA samples for validation were native DNA, if available. If not, a second whole genome amplification was performed to avoid false validation from random WGA artifacts. In the non-hypermutated patients we attempted to validate all non-silent mutations. The majority of these were validated using PCR amplification of the mutation locus, followed by sequencing on 454. Nine patients were sequenced on both Illumina and SOLiD and for those patients the initial calls observed on both platforms were considered validated. The 35 hypermutated patients accounted for 75% of the mutations in the original discovery set. These patients were subjected to a second capture, using and independent capture library preparation, followed by sequencing on an Illumina instrument. The final mutation file, Table 2A, consists of 60,313 non-silent mutations validated by these methods. "Non-silent" includes missense, nonsense, splice site or, in-frame and frameshift indels. There are an additional 7,863 primarily in non-hypermutated patients, that remain in unknown validation status, for which we will attempt further rounds of validation. Table 2A also reports 15,930 silent mutations, with 4666 remaining in unknown status.

Detection of insertion and deletion mutations in microsatellite instable tumors.
Microstatellite instability (MSI) is a hypermutator phenotype characterized by the propensity to mutate at runs of short tandem repeats (microsatellites) through insertion or deletion triggered by the disruption of the DNA mismatch repair apparatus[3]. This tumor phenotype, important because it correlates with good prognosis, is measured by assay of five tandem repeat loci. Increased rates of indel mutation leading to frameshift mutation are also observed in the coding sequence of genes harboring homopolymer runs, some of which play a key role in colorectal

adenocarcinoma[1]. We focused our analysis at the site of known homopolymer runs in selected genes (Supplementary Methods Table 5).

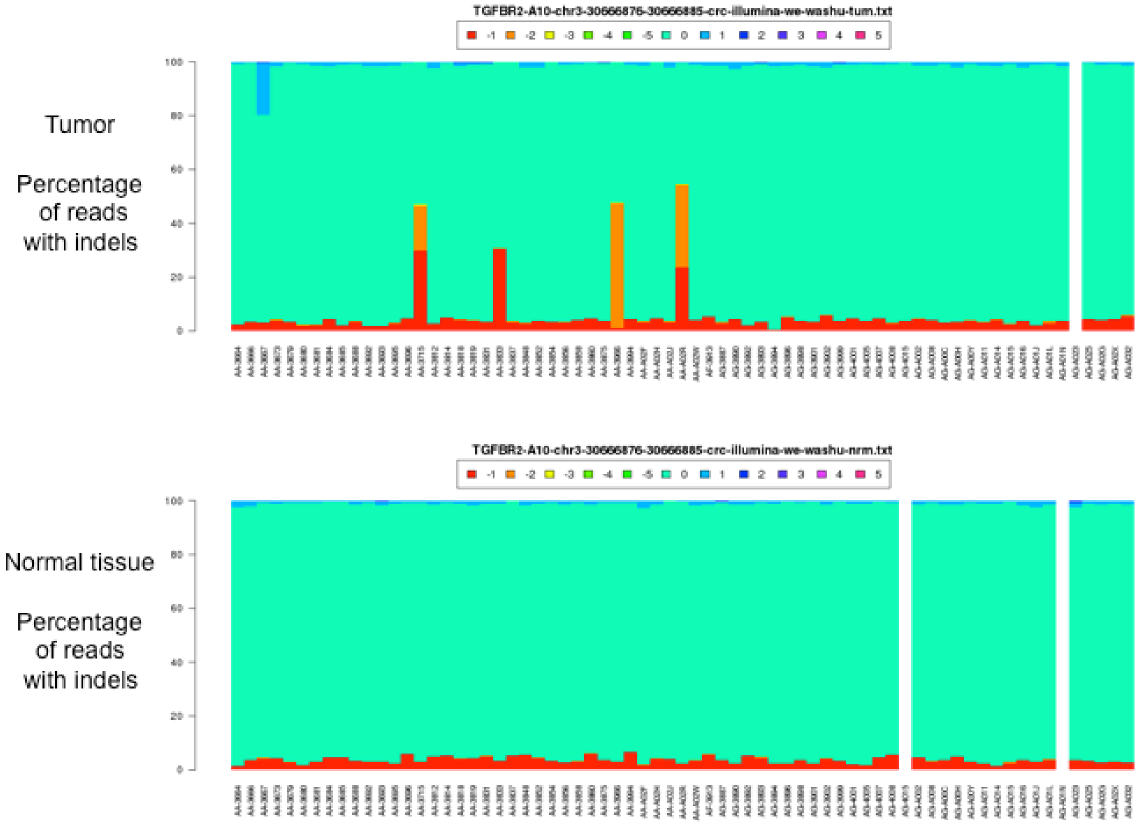
Supplementary Methods Table 5. Homopolymer runs in genes related to colorectal adenocarcinoma.



Hugo	Base	Length	Chr	Start	End
TGFBR2	A	10	3	30666876	30666885
AIM2	T	10	1	157299111	157299120
ATR	T	10	3	143757430	143757439
CASP5	T	10	11	104383251	104383260
MBD4	T	10	3	130638238	130638247
SEC63	T	10	6	108321448	108321457
SEC63	T	10	6	108321467	108321475
BLM	A	9	15	89105143	89105151
CHEK1	A	9	11	125010588	125010596
GRB14	A	9	2	165073534	165073542
PRDM2	A	9	1	13981336	13981344
RAD50	A	9	5	131959351	131959359
RHAMM	A	9	5	162850004	162850012
TCF7L2	A	9	10	114915307	114915315
WISP3	A	9	6	112496127	112496135
MLH3	T	9	14	74584357	74584365
MLH3	T	9	14	74584091	74584098
ACVR2A	A	8	2	148373511	148373518
ACVR2A	A	8	2	148400156	148400163
APAF1	A	8	12	97595338	97595345
MSH3	A	8	5	80006671	80006678
PMS2	A	8	7	5993683	5993690
PRDM2	A	8	1	13981147	13981154
MSH6	C	8	2	47884144	47884151
BAX	G	8	19	54153155	54153162
IGF2R2	G	8	6	160405478	160405485
BCL10	T	8	1	85509099	85509106
MSH2	A	7	2	47493092	47493098
AXIN2	C	7	17	60963047	60963053
FAS	T	7	10	90758688	90758694
MLH1	A	6	3	37025353	37025358
PTEN	A	6	10	89707750	89707755
PTEN	A	6	10	89710792	89710797
AXIN2	G	6	17	60963018	60963023
AXIN2	T	6	17	60963116	60963121
AXIN2	T	6	17	60956667	60956672

In our automated analysis of whole exome data, we failed to detect the predicted increase rate of frameshift mutation at these sites in patients who tested positive for MSI. Inspection of the sequencing reads at these sites revealed high levels of mutation in the tumor, but also low levels of apparent mutation in the reads from the normal. The low rate of mutation reads derived from normal tissue was most likely caused by enzyme slippage during PCR amplification[4] of the sequencing template, since this low background rate was uniform across all tumors regardless of MSI status.

If we define the mutant allele fraction (AF) as the number of reads harboring a non-reference allele divided by the total number of reads covering the locus, reads derived from normal tissue had allele fractions less than 0.01, while tumors exhibiting MSI has mutant AF greater at least 20 times higher. A histogram could be generated tallying the reference and mutated reads for each patient at a given site, as is shown in Figure A for the homopolymer run of 10 A in *TGFBR2*. To assess the rate of frameshift mutation in colorectal adenocarcinoma patients, we generated histograms, like those shown in Supplementary Methods Figure 1, for each of the genes shown in Supplementary Methods Table 5 and scored them numerically by the relative AF in tumor and normal. If the difference between the tumor and normal AF was greater than 0.2, we scored the site in the gene as mutated. The results of this analysis are shown in Supplementary Table 2.



Supplementary Methods Figure 1. Representative histogram showing indel analysis of DNA sequencing reads on an A10 homopolymer run within the TGFBR2 gene from 68 colorectal cancers (top panel) and matched normal control tissue (bottom panel). For each tumor and matched normal tissue, bars represent the fraction of read with deletions (on the bottom of each panel) or insertions (on the top of each panel). The colors of each bar vary depending on the number of nucleotides inserted or deleted (note legend above the top panel). The white bars represent missing data. Most deletions produce a -1 frameshift (red bars) but sometimes -2 deletions (orange bars) are observed. Note in the top panel that four of the tumors (with asterisk-associated bars) have much higher fractions of deletion reads than other tumors or the normal tissues. These four tumors were considered to have a frameshift mutation in one of their TGFBR2 alleles.

The frequency of mutation at a given site is dependent on the length of the homopolymer run. Runs of 10 produced the highest mutation rates whereas runs of 6 produced the lowest mutations rates.

Identification of significantly mutated genes. The ranking of genes in terms of estimated conferred selective advantage was performed using the mutation statistical analysis algorithm MutSig (v1.3 , Lawrence et al., manuscript in preparation). The MutSig algorithm works with an aggregated list of mutations across the entire patient set and estimates background mutation rates. The p and q values for a certain gene, corresponding to raw probability, p, and the probability, q, corrected for multiple testing, are determined for the mutation rate observed in that gene in relation to the background model.

MutSig uses various factors to accurately estimate the background mutation rate, taking into account the background mutation rates of different mutation categories (e.g., transitions or transversions in different sequence contexts), as well as the fact that different samples have different background mutation rates. It then uses convolutions of binomial distributions to calculate the p and values for each gene, which represents the probability that we observe by chance a certain configuration of mutations in a gene, given the background model. It also takes into account the non-synonymous to synonymous mutation ratio for each gene in order to separate out the genes with a large number of non-synonymous events compared to synonymous ones.

Copy-number and rearrangements

Identification of copy number variants. To characterize somatic copy number alterations in the tumor genome, we applied a new algorithm called BIC-seq[5] to low-coverage whole-genome sequencing data. First, we counted the reads (uniquely aligned to the reference genome with at least 46bp out of 50bp aligned) in fixed-size, non-overlapping windows along the genome. Given these bins with read counts for tumor and matched normal genomes, BIC-seq attempts to iteratively combine neighboring bins with similar copy numbers. Whether the two neighboring bins should be merged is based on Bayesian Information Criteria (BIC), a statistical criterion measuring both fitness and complexity of a statistical model. Segmentation stops when no merging of windows improves BIC, and the boundaries of the windows are reported as a final set of copy number breakpoints. Segments with copy ratio difference smaller than 0.1 (log2 scale) between tumor and normal genomes were merged in the post-processing step to avoid excessive refinement of altered regions with high read counts.

Structural variation discovery with BreakDancer. Structural Variation detection is performed with the program BreakDancer on a .bam file constructed from HiSeq sequencing of each tumor pair[6]. The first step requires a configuration file of each bam file for each tumor pair with the bam2cfg.pl perl module of the program. After the configuration file, the perl module BreakDancerMax.pl is run on the configuration file in order to call structural variants in the tumor and control files. Each tumor structural variant file is filtered with its matched normal and all possible somatic variants are filtered with a metanormal to remove any false positives.

To understand the translocations at the structural level, we PCR amplified the junction fragments using primers from regions of the two chromosomes close to the region of putative breakpoints and the DNA from this product was subjected to sequencing using the Sanger method on a capillary electrophoresis unit. Examination of the resulting sequence allowed us to define the translocation breakpoints at the nucleotide level. Results from this analysis for translocations involving NAV2-TCF7L2 and TTC28 with different partners are shown in Supplementary Figure 5.

SNP Based Copy Number Analysis. Tumor and germline-derived DNA samples were hybridized to Affymetrix SNP 6.0 arrays using manufacturers' protocols at the Genome Analysis Platform of the Broad Institute. Data are subsequently processed from the raw .CEL files using Birdseed to infer a preliminary copy-number at each probe locus. For each tumor, genome-wide copy number estimates are refined using tangent normalization, in which tumor signal intensities are divided by signal intensities from the linear combination of all normal samples that are most similar to the tumor (to be described in greater detail in Getz et al, in preparation).

This linear combination of normal samples tends to match the noise profile of the tumor better than any set of individual normal samples, thereby reducing the contribution of noise to the final copy-number profile. The individual copy-number estimates then undergo segmentation using Circular Binary Segmentation as described previously. As part of this process of copy-number assessment and segmentation, we remove regions corresponding to germline copy-number alterations by elimination of regions of copy-number alteration identified from either previously annotated copy-number variant filters created using the TCGA germline samples for the ovarian cancer analysis or from additional removal of regions identified in the germline genomes of samples from this collection.

All samples following processing are analyzed by several quality control metrics. Both tumor and normal samples are screened for noise, as evidenced by excessive variation between successive probes or an excessive segmentation count. In addition, normal samples are screened for DNA quality by the Affymetrix FQC probes on the SNP 6.0 array as well as by the genotyping call rate. Among the samples in this study entering into the copy-number pipeline, 7% of tumors and 28% of germline samples failed to meet all quality-control metrics. The higher percentage of germline samples failing this metric was attributed to some variation in DNA extraction methods used for several batches of germline DNA.

As described previously, the segmented copy number profiles for ovarian carcinoma and matched control DNAs were first analyzed using *Ziggurat Deconstruction* to determine the length and amplitude of inferred copy number changes underlying each segmented copy number profile[7,8]. Events are subsequently categorized into focal copy number events much smaller than a chromosome arm, and broad copy number events that span a chromosome arm or entire chromosome for separate analysis of the two classes of events. As with the TCGA ovarian cancer analysis, length threshold of 50% of a chromosome arm was used to distinguish between broad and focal events. Segments were filtered using an amplitude threshold at a copy-difference of a log2 copy-number ratio of 0.3 with amplification intensities capped at a value of 1.5 to avoid hypersegmentation due to variation of the dynamic range of probes on the SNP arrays. Analysis of broad copy number changes was performed as previously described[9]. Similarly, focal copy number changes in the 186 colon adenocarcinoma samples, 71 rectal adenocarcinoma samples and the composite dataset were analyzed using the GISTIC methodology[9]. All focal amplifications and deletions identified by GISTIC were subject to visual inspection, and the events representing likely segmentation artifacts or germline variants (amplification peaks at intergenic regions on 2q21.2, 7q34 and 17p11.1) were removed from analysis. For the GISTIC analysis, those events with false discovery rates (FDR) <0.05 were included for discussion in the text of the manuscript and downstream analyses. For completeness, Supplementary Table 4 lists all GISTIC amplification and deletion peaks to an FDR of 0.25. However, with the peaks at lower statistical significance, GISTIC was often less able to resolve those peaks to more focused sets of target genes.

Microarray Expression Profiling

RNA labeling, array hybridization and data processing. Gene expression profiling was performed as described previously[9]. Briefly, 2 ug of total RNA of sample (n=220) and Stratagene Universal Human Reference were amplified and labeled using Agilent’s Low RNA Input Linear Amplification Kit. Sample and reference were co-hybridized on a Custom Agilent 244K Gene Expression Microarray (AMDID019760). The expression data was Lowess normalized and the ratio of the Cy5 channel (sample) and Cy3 channel (reference) were log2 transformed to create gene expression values for 23,199 probesets. Probesets without gene annotations and genes with missing data in ≥ 20% of the samples were removed, resulting in 13,994 genes available for further analysis. Missing values in the remaining genes were imputed with the mean value across all samples. PCA (JMP Genomics, v.4.0) analysis indicated that the source of the RNA (BCR) was responsible for 23% of the variance of the microarray data, which was normalized out using JMP Genomics' Batch Correction procedure

Consensus and Hierarchical Clustering. The normalized gene expression dataset of 220 colorectal cancers (152 colon and 68 rectum) was filtered to include 1,558 consistent, but variably expressed genes (MAD > 0.564). Consensus clustering[10] using self organizing maps identified between 2-4 robust clusters. A nearest centroid-based classifier (CLaNC) was used to identify signature genes for each of the clusters[11]. Grouping the samples into 3 subtypes resulted in the identification of gene signature containing 1020 genes (340 genes per class), which had the lowest cross validation and prediction error out of the 2, 3 and 4 class classifiers Hierarchical clustering of the 220 colorectal cancers across the 1020 signature genes demonstrates that the samples group into 3 distinctive groupings (Supplementary Figure 2).

RNA-Seq Methods

Total RNA for each sample was converted into a library of template molecules for sequencing on the Illumina Cluster Station and Genome Analyzer according to the protocol for the Illumina mRNA Sample preparation kit (Part#1004898, Rev A: Illumina, San Diego, CA). Briefly, poly-A mRNA was purified from total RNA (2 µg) using poly-T oligo-attached magnetic beads. The mRNA was then fragmented and the first strand of cDNA was synthesized from the cleaved RNA fragments using reverse transcriptase and random primers. Following the synthesis of the second strand of cDNA, end repair was performed on overhangs using T4 DNA polymerase and Klenow DNA polymerase, followed by ligation of sequencing Adapters to the ends of the DNA fragments. The cDNA fragments were purified using a gel run at 80 V for approximately 3 hours until the Orange G dye band reached the bottom of the gel. The gel was stained with SYBR green to visualize the DNA band. A band at 350 – 450 bp was excised vertically from the gel, which was then dissolved at room temperature using a QIAquick Gel Extraction Kit (Qiagen, Valencia, CA). The purified cDNA templates were enriched for 15 cycles of PCR amplification and validated using a BioAnalyzer to assess size, purity and concentration of the purified cDNA libraries. The cDNA libraries were placed on an Illumina Cluster Station for single end cluster generation according to the protocol outlined in the Illumina Genome Analysis User Guide (Part# 11251649, RevA). The template cDNA libraries (1.5 µg) were hybridized to a flow cell, amplified and linearized and denatured to create a flow cell with ssDNA ready for sequencing. Each flow cell was sequenced on an Illumina GAIIX Genome Analyzer. Each sample underwent a single lane of sequencing using single end sequencing for 76 cycles according to the protocol outlined in the Illumina Genome Analysis User Guide (Part# 11251649, RevA). After completion of the 76 cycle sequencing run, the raw sequence data entered the UNC RNAseq Workflow.

RNA-Seq Data Processing Workflow (BWA to Transcriptome)

Step 1: Construct database of reference transcript sequences, composite gene models, composite exons, and splice junctions. The reference transcript set is based on the hg19 UCSC Gene standard track (December 2009 version), which is a publically-available sequence set that can be downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/). From this set, only sequences mapped to canonical human chromosomes (chr1-22, X, Y, and M) were retained. For each selected transcript, the nucleotide sequence, association to Entrez/LocusLink genes (if known), and CDS range (if known) were extracted from the UCSC database tables. Additionally, a pairwise alignment of each transcript against the hg19 genome was provided by Mark Diehkans from UCSC. The reference transcript set contains 73,671 human transcript sequences representing 20,532 human genes. 67,434 transcript sequences are given gene assignments according to the UCSC database tables; the remaining 6,237 sequences in the reference transcript set are not currently associated with an Entrez/LocusLink gene. No transcript is associated with more than one Entrez/LocusLink gene. For each gene represented by one or more transcripts in this set, a composite gene model was generated by merging all overlapping exons (as defined by the genomic mapping) from each associated reference transcript. Thus, each composite gene model is essentially the union of all associated reference transcripts. Each gene model is linked to a specific Entrez/LocusLink gene identifier. A library of composite exons was then established by collecting each contiguous genomic segment from each gene model. Note that because the gene model construction process involves merging overlapping exons from different transcripts, the resulting composite exons may or may not be observed in reference transcript sequences. This library contains 239,886 unique composite exons, each of which is defined by its genomic range. A library of all known splice junctions was established by cataloging all junctions observed in the reference transcript sequences. Inclusion of a splice junction in this library indicates that it has been observed in (at least) one reference transcript. However, not all splice junctions in this library will be observed in the composite gene models. This library contains 249,775 unique splice junctions, where each junction is defined by the last position of exon N and the first position of exon N+1.

Step 2: Prepare raw sequencing reads for alignment. The "illumina2srf" tool from the DNA Sequence Read Toolkit (http://sourceforge.net/projects/sequenceread/) is used to convert the raw sequencing data from the vendor-specific format to standard SRF (sequence read format), which is a preferred compressed format for storing sequencing reads. Before data processing begins, the "srf2fastq" tool from the Staden Package (http://staden.sourceforge.net/) is utilized to convert the data from SRF to the FASTQ format, using the "-C" parameter to simultaneously filter poor quality reads. In some cases, the insert size of the sequenced fragment is shorter than the read length, resulting in part or all of the adapter (primer) sequence being incorporated into the output read. Since any read containing adapter sequence is highly unlikely to align to a biological reference database, all reads are further processed to trim any adapter segments from the sequences. The criteria for identifying adapter sequence within a read are as follows: (a) the read contains an exact match with the adapter sequence, (b) the sequence match begins at the first base of the adapter, (c) the sequence match continues until either the end of the adapter or the end of the read, and (d) the sequence match is at least 5 bases in length. If all criteria are met, the read is trimmed starting at the first base of the adapter sequence match.

Step 3: Align sequencing reads against reference transcript database. A pre-processed lane of sequencing reads is then submitted to the BWA algorithm (http://bio-bwa.sourceforge.net/) for alignment against the reference transcript database using default parameters. The resulting SAM file is converted to BAM format using Picard tools (http://picard.sourceforge.net/).

Step 4: Calculate quantification at the transcript level. The reads aligned to the reference transcript sequences are used to determine transcript level quantification. Three quantification values are reported: raw read counts, coverage, and RPKM. Raw read counts is simply the number of reads aligned to a given reference transcript. Raw read counts are normalized by transcript length to give coverage. For a given TranscriptX, coverage is calculated by: total bases aligned to TranscriptX / length of TranscriptX. The "total bases aligned" will typically be equal to aligned reads × read length. However, in cases where reads have been trimmed due to adapter contamination, this value may be slightly lower. Raw read counts are normalized by both transcript length and overall lane yield to give RPKM. For a given TranscriptX, RPKM is calculated by: (raw read counts × 10^9) / (total reads × length of TranscriptX). For this calculation, "total reads" is the lane yield after removing poor quality reads.

Step 5: Calculate quantification at the gene level. Quantification at the gene level is determined by collapsing the transcript level quantifications onto their associated gene loci. Raw read counts, coverage, and RPKM are reported for each gene. Raw read counts for a given GeneX is the sum of reads aligned to all transcripts associated with that gene. Coverage for a given GeneX is determined by: sum of bases aligned to all transcripts associated with GeneX / length of GeneX. RPKM for a given GeneX is calculated by: (raw read counts × 10^9) / (total reads × length of GeneX). "Total reads" is the lane yield after removing poor quality reads. In both the coverage and RPKM calculations, the length of GeneX is defined as the median length of all transcripts associated with GeneX.

Step 6: Calculate quantification at the exon level. In order to carry quantification for sequence features defined by their genomic locus (i.e. exons and splice junctions), the aligned reads must first be translated from transcript coordinates to genomic coordinates. The pre-established pairwise mapping between each reference transcript and the hg19 genome allows for a straightforward conversion between these two coordinate systems. These converted aligned reads are then used to determine exon level quantification. Raw base counts, coverage, and RPKM are reported for each entry in the library of composite exons. The raw base counts for a given ExonX is the total number of bases aligned to that genomic segment. Raw base counts are used instead of raw read counts because in many cases only a portion of a read will align to a given exon. For a given ExonX, coverage is calculated by: raw base counts / exon length. RPKM for a given ExonX is determined by: (raw base counts / median read length) × 10^9) / (total reads × exon length). Except in cases where the raw reads have undergone extensive adapter trimming, the median read length will be equal to the experimental read length for that lane. "Total reads" is the lane yield after removing poor quality reads.

Step 7: Calculate quantification at the splice junction level. Quantification at the splice junction level is also calculated based on the aligned reads converted to genomic coordinates. The only reported value for this level is raw read counts, which is determined by the number of reads that cross a particular junction. Because a splice junction has an effective length of zero, the coverage and RPKM calculations do not apply.

Step 8: Generate visualization of coverage. To facilitate visualization of read coverage across genomic segments of interest, the quantification data is converted to the UCSC bigWig format to enable viewing in the UCSC Genome Browser with down to single-base resolution.

Step 9: Evaluate QC metrics. Each lane is evaluated according to a variety of both pre- and post-alignment QC measures in order to determine whether the data set in question falls within expected quality thresholds. The following metrics are assessed on a per-lane basis:

- Total read counts
- Reads passing filter

- Percent of reads that are unique
- Base quality per cycle
- Nucleotide distribution per cycle
- Percent of reads requiring adapter trimming
- Mean effective read length after adapter trimming
- Percent of reads aligning to human reference transcripts
- Percent of reads aligning to human rRNA
- Percent of reads aligning to human miRNA
- Percent of reads aligning to human mtDNA
- Percent of reads aligning to the human genome
- Percent of reads aligning to other genomes (mouse, rat, fruit fly, Arabidopsis, yeast)
- Percent of reads aligning to viral genomes
- Average coverage per human gene
- Average coverage across the length of reference transcripts

RNAseq determination of PI3KCA sequence variants

The subset of samples that was sequenced on the SOLiD platform had poor coverage for PIK3CA. We used RNA-Seq data to call mutations in PIK3CA in these samples. For each sequenced sample a pileup file was generated from the aligned .bam file using the Samtools (http://samtools.sourceforge.net) pileup function. Sequence variants at a given chromosomal location were considered valid if the variant base was present in ≥ 30% of the reads at the location. The identified mutations are listed in Supplementary Table 2.

DNA Methylation Profiling

Array-based DNA methylation assay. We used the Illumina Infinium DNA methylation platform (HumanMethylation27 BeadChip) (Illumina, San Diego, CA) to obtain gene promoter DNA methylation profiles of 167 TCGA colon adenocarcinoma samples and 37 adjacent non-tumor colonic tissue samples (batches 28-30, 33, 36, 41, 45, 66) and 69 TCGA rectal adenocarcinoma samples and five adjacent non-tumor rectal tissue samples (batches 42 and 46). The Infinium HumanMethylation27 panel targets 27,578 CpG sites located in proximity to the transcription start sites of 14,475 consensus coding sequencing (CCDS) in the NCBI Database (Genome Build 36). The assay probe sequences and information on each interrogated CpG site on the Infinium HumanMethylation27 BeadChip can be found in the MAGE-TAB ADF (Array Design Format) file deposited on the TCGA Data Portal. We performed bisulfite conversion on 1 µg of genomic DNA from each sample using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA) according to the manufacturer's instructions. We assessed the amount of bisulfite converted DNA and completeness of bisulfite conversion using a panel of MethyLight-based quality control (QC) reactions as previously described[12] (10). All the TCGA samples passed our QC tests and entered the Infinium DNA methylation assay pipeline.

Bisulfite-converted DNA was whole genome amplified (WGA) and enzymatically fragmented prior to hybridization to BeadChip arrays. The amplified and fragmented DNA molecules anneal to a locus-specific DNA oligomers (50 mers) covalently attached to a specific bead type. Each interrogated CpG locus can hybridize to methylated (CpG) or unmethylated (TpG) oligo bead types. DNA methylation-specific primer annealing is followed by single-base extension using labeled nucleotides [cy5 (red) or cy5 (green)]. Both methylated (M) and unmethylated (U) bead types for a specific CpG locus incorporate the same labeled nucleotide, as determined by the base immediately preceding the cytosine being interrogated by the assay, and are subsequently detected in a single color channel. Fluorescence intensities of the M and U bead types for each CpG locus were measured using the Illumina BeadArray Reader. The mean signal intensities for replicate M and U probes for each CpG locus were extracted from Illumina GenomeStudio software. The level of DNA methylation at each CpG locus is scored as beta (β) value calculated as (M/(M+U)), ranging from 0 to 1, with values close to 0 indicating low levels of DNA methylation and beta values close to 1 indicating high levels of DNA methylation.

The detection *P* values provide an indication of the quality of DNA methylation measurement and are calculated as previously described[13]. We determined that data points with a detection *P* value >0.05 are not significantly different from background measurements, and therefore were masked as “NA” in the Level 2 and Level 3 data packages. All Infinium DNA methylation data were packaged and deposited onto the TCGA Data Portal web site (http://tcga-data.nci.nih.gov/tcga/).

TCGA data packages. The data levels and the files contained in each data level package are described below and are present on the TCGA Data Portal website (http://tcga-data.nci.nih.gov/tcga/). Please note that with continuing updates of genomic databases, data archive revisions become available at the TCGA Data Portal.

LEVEL 1: Level 1 data contain the non-background corrected signal intensities of the M and U probes and the mean negative control cy5 (red) and cy3 (green) signal intensities. A detection *P* value for each data point, the number of replicate beads for M and U probes as well as the standard error of M, U, and control probe signal intensities are also provided. It is important to note that for some CpG targets, both M and U measurements will be cy3, and for others both will be cy5. To resolve ambiguities regarding this subtlety of the Infinium DNA Methylation assay, we have labeled the cy3 and cy5 values deposited to the DCC as “Methylated Signal Intensity” and “Unmethylated Signal Intensity”. The information of which dye is used for each CpG locus is supplied in the MAGE-TAB ADF file deposited in the DCC.

LEVEL 2: Level 2 data files contain the β-value calculations for each probe and sample. Data points with detection *P* values >0.05 were not considered to be significantly different from background, and were masked as “NA”.

LEVEL 3: Level 3 data contain β-value calculations, HUGO gene symbol, chromosome number and genomic coordinate for each targeted CpG site on the array. In addition, we masked data points with "NA" from the probes that 1) contain known single nucleotide polymorphisms (SNPs) after comparison to the dbSNP database (Build 130), 2) contain repetitive sequence elements that cover the targeted CpG locus in each 50 bp probe sequence, 3) are not uniquely aligned to the human genome (NCBI build 36.1) at 20 nucleotides at the 3' terminus of the probe sequence, 4) span known regions of small insertions and deletions (indels) in the human genome (dbSNP build 130).

The following data archives were used for the analyses described in this manuscript.

Colon adenocarcinoma samples and normal-adjacent colon tissues:
Batch 28: jhu-usc.edu_COAD.HumanMethylation27.Level_3.1.2.0
Batch 29: jhu-usc.edu_COAD.HumanMethylation27.Level_3.2.2.0
Batch 30: jhu-usc.edu_COAD.HumanMethylation27.Level_3.3.1.0
Batch 33: jhu-usc.edu_COAD.HumanMethylation27.Level_3.4.1.0
Batch 36: jhu-usc.edu_COAD.HumanMethylation27.Level_3.5.2.0
Batch 41: jhu-usc.edu_COAD.HumanMethylation27.Level_3.6.0.0
Batch 45: jhu-usc.edu_COAD.HumanMethylation27.Level_3.7.0.0
Batch 66: jhu-usc.edu_COAD.HumanMethylation27.Level_3.8.0.0
jhu-usc.edu_COAD.HumanMethylation27.mage-tab.1.15.0

Rectal adenocarcinoma samples and normal-adjacent rectum tissues:
Batch 28: jhu-usc.edu_READ.HumanMethylation27.Level_3.1.2.0
Batch 29: jhu-usc.edu_READ.HumanMethylation27.Level_3.2.1.0
Batch 30: jhu-usc.edu_READ.HumanMethylation27.Level_3.3.1.0
Batch 33: jhu-usc.edu_READ.HumanMethylation27.Level_3.4.1.0
Batch 36: jhu-usc.edu_READ.HumanMethylation27.Level_3.5.1.0
Batch 46: jhu-usc.edu_READ.HumanMethylation27.Level_3.6.0.0
jhu-usc.edu_READ.HumanMethylation27.mage-tab.1.13.0

See Supplementary Table 1 for a complete map between samples and archives. It should be noted that the READ samples for batches 28, 29, 30, 33 and 36 were reorganized in the TCGA Data Portal as belonging to Batch 42. This information is also provided in the Description file for READ archives 1.2.0, 2.1.0, 3.1.0, 4.1.0 and 5.1.0.

Unsupervised clustering analysis of DNA methylation data. Statistical analysis and data visualization were carried out using the R/Biocoductor software packages (http://www.bioconductor.org). We used recursively partitioned mixture model (RPMM) for the identification of colorectal tumor subgroups based on the Illumina Infinium DNA methylation data. RPMM is a model-based unsupervised clustering approach well-suited for beta-distributed DNA methylation measurements which lie between 0 and 1, and implemented as *RPMM* R/Bioconductor package[14]. We first removed probes which contain any “NA”-masked data points and probes that are designed for the sequences on X and Y chromosomes. We then performed RPMM clustering on 2,758 probes (10% of all the original probe set) that showed the most variable DNA methylation levels based on standard deviations across the colorectal tumor panel. A fanny algorithm (a fuzzy clustering algorithm) was used for initialization and level-weighted version of Bayesian information criterion (BIC) as a split criterion for an existing cluster as implemented in the *RPMM* package. The DNA methylation β-values were represented graphically using a heatmap, generated by the R package *Heatplus*. Ordering of the samples within a RPMM class in the heatmaps was obtained by using the *seriatio*n R package.

Integration of DNA methylation and gene expression data. We used Level 3 DNA methylation data on 23,094 CpG sites covering 13,387 genes, and Level 3 lowess normalized gene expression data set on 17,814 genes generated on the UNC-Agilent 244K custom gene expression array platform. These two data sets were merged by gene symbols. A set of 21,325 Infinium DNA methylation probes associated with 12,329 genes has matched gene expression data. We have 238 samples including 223 tumor and 15 adjacent-normal tissue samples that have both DNA methylation and expression data.

Identification of epigenetically silenced genes in CRC. We determined candidate epigenetically silenced genes in CRC using the method previously developed[9]. Briefly, we considered the four criteria each with a relaxed and a stringent thresholds: 1) The mean DNA methylation β-value in non-tumor adjacent colonic tissue < 0.5 (relaxed) and < 0.4 (stringent); 2) The difference in DNA methylation β-value between the 90th percentile tumor and mean adjacent-normal > 0.1 (relaxed) and > 0.3 (stringent); 3) The fold expression change between mean adjacent-normal and mean of the 10% of tumor samples with the highest DNA methylation > 1.5 (relaxed) and > 3 (stringent); 4) Spearman's correlation coefficient between DNA methylation and gene expression calculated jointly across 223 tumor and 15 adjacent-normal tissue samples < –0.2 (relaxed) and < –0.3 (stringent). We required candidate epigenetically silenced genes to pass all four relaxed thresholds, and at least three out of four more stringent thresholds. If there were multiple probes for the same gene, the probes with the highest absolute Spearman's Rho was retained for that gene. A complete list of the 355 genes is shown in Supplementary Table 12, ranked by descending absolute Spearman's Rho.

Identification of MLH1 epigenetically silenced cases. We assessed the *MLH1* DNA methylation status in each sample based on the probe (cg00893636) located in the bidirectional *MLH1*/*EPM2AIP1* promoter CpG island and closest to the current RefSeq *MLH1* transcription start sites. DNA methylation at this site showed a strong inverse relationship with *MLH1* expression (Spearman's Rho = −0.32). We re-scaled the *MLH1* expression data between 0 and 1 (as is presented for DNA methylation β-values). We then performed K-means clustering (K=2) on the two-dimensional space of DNA methylation and expression data to classify the epigenetically silenced group and non-epigenetically silenced group of samples.

miRNA-Seq

Library construction and sequencing. RNA samples, including controls, are arrayed into 96-well plates and a subset of 12 samples using an Agilent Bioanalyzer RNA nanochip.

Two micrograms of total RNA are separated into mRNA and miRNA fractions. Briefly, total RNA is mixed with oligo(dT) Microbeads and loaded into a 96-well MACS column that is placed on a MultiMACS separator (Miltenyi Biotec, Germany). The separator's strong magnetic field allows beads to be captured during washes. From the flow-through, small RNAs, including miRNAs, are recovered by ethanol precipitation. Quality is checked for a subset of 12 samples using an Agilent Bioanalyzer RNA nanochip.

miRNA-Seq libraries are constructed using a British Columbia Genome Sciences Centre (BCGSC) plate-based protocol. The RNA recovered from the flow-through is arrayed onto a 96-well plate with a positive control RNA sample. Negative controls are added at three stages; elution buffer is added to two wells when the total RNA is loaded into the plate, water to one well just before ligating the 3' adapter, and water to another well just before PCR. A 3' adapter is ligated using a truncated T4 RNA ligase2 (NEB Canada, cat# M0242L) with an incubation of 1 hour at 22°C. This adapter is adenylated, single-strand DNA with the sequence 5' /5rApp/ ATCTCGTATGCCGTCTTCTGCTTGT /3ddC/, which selectively ligates miRNAs. An RNA 5' adapter is then added, with T4 RNA ligase (Ambion USA, Cat#AM2141) and ATP, and is incubated at 37°C for 1 hour. The sequence of the single strand RNA adapter is 5'GUUCAGAGUUCUACAGUCCGACGAUCUGGUCAA3'.

When ligation is complete, 1st strand cDNA is synthesized using Superscript II Reverse Transcriptase (Invitrogen, cat#18064 014) and RT primer (5'-CAAGCAGAAGACGGCATACGAGAT-3’). This is the template for the final library PCR, into which we introduce index sequences to enable libraries to be identified from a sequenced pool of libraries. Briefly, a PCR brew mix is made with the 3’ PCR primer (5’-CAAGCAGAAGACGGCATACGAGAT-3’), Phusion Hot Start High Fidelity DNA polymerase (NEB Canada, cat# F-540L), buffer, dNTPs and DMSO. The mix is distributed evenly into a new 96-well plate. A Biomek FX (Beckman Coulter, USA) is used to transfer the PCR template (1st strand cDNA) and indexed 5’ PCR primers into the brew mix plate. Each indexed 5’ PCR primer, 5'-AATGATACGGCGACCACCGACAGNNNNNNGTTCAGAGTTCTACAGTCCGA-3’, contains a unique six-nucleotide ‘index’ (shown here as N’s), and is added to each row of the 96-well PCR brew plate, resulting in 8 indexed primers being used in each column of the plate. PCR is run at 98°C for 30 sec, followed by 15 cycles of 98°C for 15 sec, 62°C for 30 sec and 72°C for 15 sec, and finally a five min incubation at 72^{o}C. Upon PCR completion, quality is checked for a subset of 12 samples using an Agilent Bioanalyzer DNA1000 chip, and is reported as an RNA integrity number (RIN). PCR products are pooled by column (8 indices per pool) and are size selected to remove larger cDNA fragments and adapter contaminants, using a BCGSC-developed 96-channel automated size selection robot. After size selection, each pool is ethanol precipitated, quality checked using an Agilent Bioanalyzer DNA1000 chip and quantified using a Qubit fluorometer (Invitrogen, cat# Q32854). Each pool of 8 libraries is diluted to a target concentration for cluster generation and then is loaded onto a single lane of an Illumina GAiix flow cell. Clusters are generated, and a 31bp main read and a 7bp index read are sequenced.

Preprocessing, alignment and annotation. Raw sequence data are preprocessed before being analyzed. Briefly, after an initial QC, the data are separated into individual samples by matching the index sequence to the start of each through-index read. The index sequence and the adapter sequence at the end of each read are removed, and the reads for each sample are aligned to the NCBI36 reference genome using BWA[15]. Below, we describe these steps in more detail.

In a routine QC stage, 20% of raw sequences from each 8-way pooled lane are checked for the abundance of reads in each indexed sample, and for the proportion of reads that are possibly from adapter dimers (i.e. a 5’ adapter joined to a 3’ adapter with no intervening biological sequence) or from miRNAs of different species. Sequencing error is estimated independently by a method originally developed for SAGE[16].

Libraries that pass this QC stage are preprocessed for alignment. The size-selected miRNAs vary somewhat in length and tend to be shorter (~20 bp) than the read length. Given this, each read sequence extends some distance into the 3' sequencing adapter. Because this non-biological sequence can interfere with aligning the read to the reference genome, 3’ adaptor sequence is identified and removed (trimmed) from a read. The adapter-trimming algorithm identifies as long an adapter sequence as possible, allowing a number of mismatches that depends on the adapter length found. A typical sequencing run yields several million reads; using only the first (5’) 15 bases of the 3’ adapter in trimming makes processing efficient, while minimizing the chance that an miRNA read will match the adapter sequence.
The algorithm first determines whether a read sequence should be discarded as an adapter dimer by checking whether the 3’ adapter sequence occurs at the start of the read. For reads passing this stage, the algorithm then tries to identify an exact 15-bp match anywhere within the read sequence. If it cannot, it then retries, starting from the 3' end, and allowing up to 2 mismatches. If the full 15bp is not found, decreasing lengths of adapter are checked, down to

the first 8 bases, allowing one mismatch. If a match is still not found, from 7 bases down to 1 base is checked, with an exact match required. Finally, the algorithm will trim 1 base off the 3' end of a read if it happens to match the first base of the adapter. This is based on two considerations. First, it is preferable to get a perfect alignment than an alignment that has a potential one-base mismatch. Second, if only 1 base of adapter was found in the read sequence, the read is likely too long to be from a miRNA and the effect of the trimming on its alignment would not affect this sample's overall miRNA profiling result. After each read has been processed, a summary report is generated containing the number of reads at each read length. Because the shortest mature miRNA in miRBase v13 is 15 bp, any trimmed read that is shorter than 15bp is discarded; remaining reads are submitted for alignment to the reference genome. Alignment(s) for each read are checked with a series of three filters. A read with more than 3 alignments is discarded as too ambiguous. For TCGA quantification reports, only perfect alignments with no mismatches are used. Based on comparing expression profiles of test libraries (data not shown), reads that fail the Illumina basecalling chastity filter are retained, while reads that have soft-clipped BWA CIGAR strings are discarded.

For reads retained after filtering, each coordinate for each read alignment is annotated using the reference databases (Supplementary Methods Table 4), and requiring a minimum 3-bp overlap between the alignment and an annotation. In annotating reads we address two potential issues. First, a single read alignment can overlap feature annotations of different types; second, a read can have up to three alignment locations, and each alignment location can overlap a different type of feature annotation. By considering heuristically determined priorities (Supplementary Methods Table 4), we resolve the first issue by giving each alignment a single annotation. We resolve the second by collapsing multiple annotations to a single annotation, as follows.

If a read has more than one alignment location, and the annotations for these are different, we use the priorities from Supplementary Methods Table 4 to assign a single annotation to the read, as long as only one alignment is to a miRNA. When there are multiple alignments to different miRNAs, the read is flagged as cross-mapped, and all of its miRNA annotations are preserved, while all of its non-miRNA annotations are discarded. This ensures that all annotation information about ambiguously mapped miRNAs is retained, and allows annotation ambiguity to be addressed in downstream analyses. Note that we consider miRNAs to be cross-mapped only if they map to different miRNAs, not to functionally identical miRNAs that are expressed from different locations in the genome. Such cases are indicated by miRNA names. A miRBase name can have up to 4 separate sections separated by "-", e.g. hsa-mir-26a-1. A difference in the final (e.g. '-1') section denotes functionally equivalent miRNAs expressed from different regions of the genome, and we consider only the first 3 sections (e.g. 'hsa-mir-26a') when comparing names. As long as a read maps to multiple miRNAs for which the first 3 sections of the name are identical (e.g. hsa-mir-26a-1 and hsa-mir-26a-2), it is treated as if it maps to only one miRNA, and is not flagged as cross-mapped.

From the profiling results for a tumor type, for a minimum of approximately 100 samples, we identify the depth of sequencing required to detect the miRNAs that are expressed in a sample by considering a graph of the number of miRNAs detected in a sample as a function of the number of reads aligned to miRNAs. For any sequencing run that fails to meet this threshold, we sequence the sample again to achieve at least the tumor-specific minimum number of miRNA-aligned reads.

Finally, for each sample, the reads that correspond to particular miRNAs are summed and normalized to a million miRNA-aligned reads to generate the quantification files that are submitted to the DCC. Quantification files include information on variable 5’ and 3’ read alignment locations, which can reflect isoforms, adapter trimming and RNA degradation.

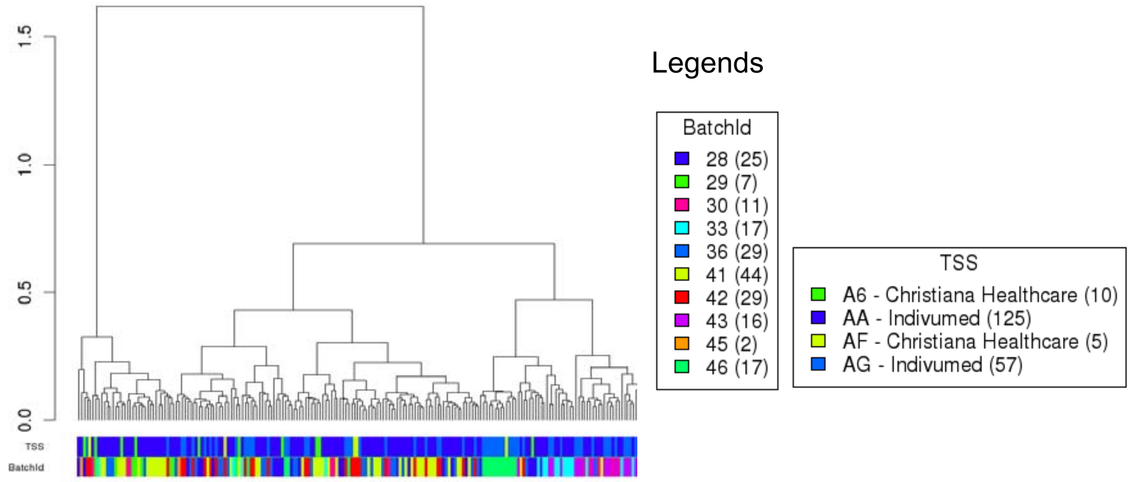
Consensus clustering of miRNAs. Normalized abundance profiles for 255 COAD and READ samples were clustered using 200 iterations of NMF v0.5.2[17], for 3 to 10 clusters. Inspection of cophenetic correlation coefficients and overall average silhouette width[18] for each clustering result suggested that 3 clusters was a preferred result, with 5 and 8 clusters also of interest (data not shown). A 3-cluster NMF result was then generated using 1000 iterations, and a silhouette plot was generated for this result, using R 2.13.0.

Batch effects analysis

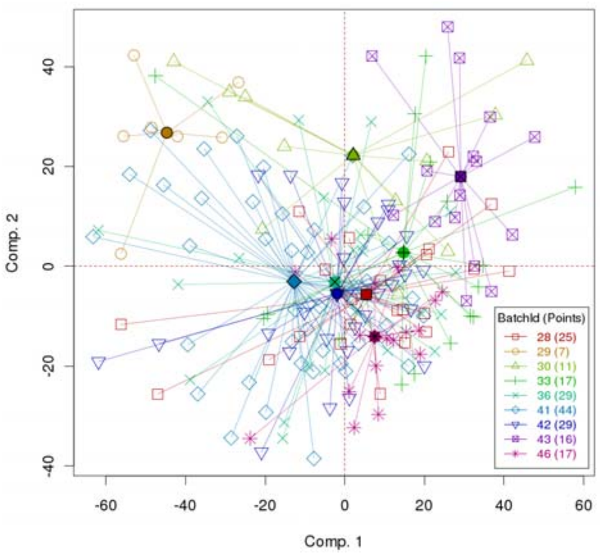
We used hierarchical clustering and Principal Components Analysis (PCA) to assess batch effects in the colorectal data sets. Five different data sets were analyzed: mRNA expression (Agilent G4502A microarray), mRNA expression (RNA-seq Illumina GA), miRNA expression (RNA-seq Illumina GA), DNA methylation (Infinium HM27 microarray), and SNPs (GW SNP 6). All of the data sets were at TCGA level 3, since that’s the level on which most of the analyses in the paper are based. We assessed batch effects with respect to two variables; batch ID and Tissue Source Site (TSS).

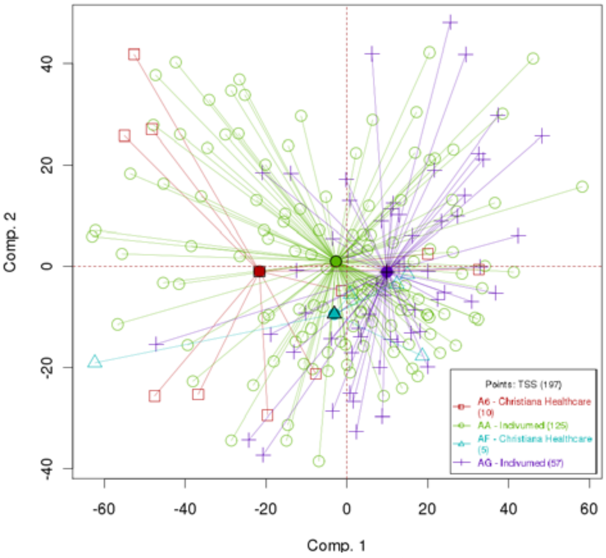
For hierarchical clustering, we used the average linkage algorithm with 1 minus the Pearson correlation coefficient as the dissimilarity measure. We clustered the samples and then annotated them with colored bars at the bottom. Each color corresponded to a batch ID or a TSS. For PCA, we plotted the first four principal components, but only plots of the first two components are shown here. To make it easier to assess batch effects, we enhanced the traditional PCA plot with centroids. Points representing samples with the same batch ID (or TSS) were connected to the batch centroid by lines. The centroids were computed by taking the mean across all samples in the batch. That procedure produced a visual representation of the relationships among batch centroids in relation to the scatter within batches. For the CRC data sets, if both hierarchical clustering and PCA suggested a batch effect, we studied that effect more closely (see batch 29 in the Conclusions section, for example). The results for the five data sets follow.

mRNA Expression (Agilent G4502A microarray). Supplementary Methods Figures 2-4 show clustering and PCA plots for the Agilent G4502A mRNA expression platform. We noticed that batch 29 stood out from the others (yellow cluster in top left of Suppl. Methods Fig. 3). Both clustering and PCA plots flagged batch 29 as distinct from the others. Overall, the samples from Christiana Healthcare, for instance (red cluster in Suppl. Methods Fig. 4.), mixed in with samples from Indivumed (green cluster in Suppl. Methods Fig. 4) and didn’t stand out in its own cluster in Suppl. Methods Fig. 2. Batch 30 seemed somewhat distinct in Suppl. Methods Fig. 3 but didn’t cluster separately in Suppl. Methods Fig. 2. But batch 29 consistently stood out when the other data types were analyzed. More results and discussion of batch 29 follow.



Suppl. Methods Fig. 2. Hierarchical clustering plot for mRNA expression (Agilent microarray)

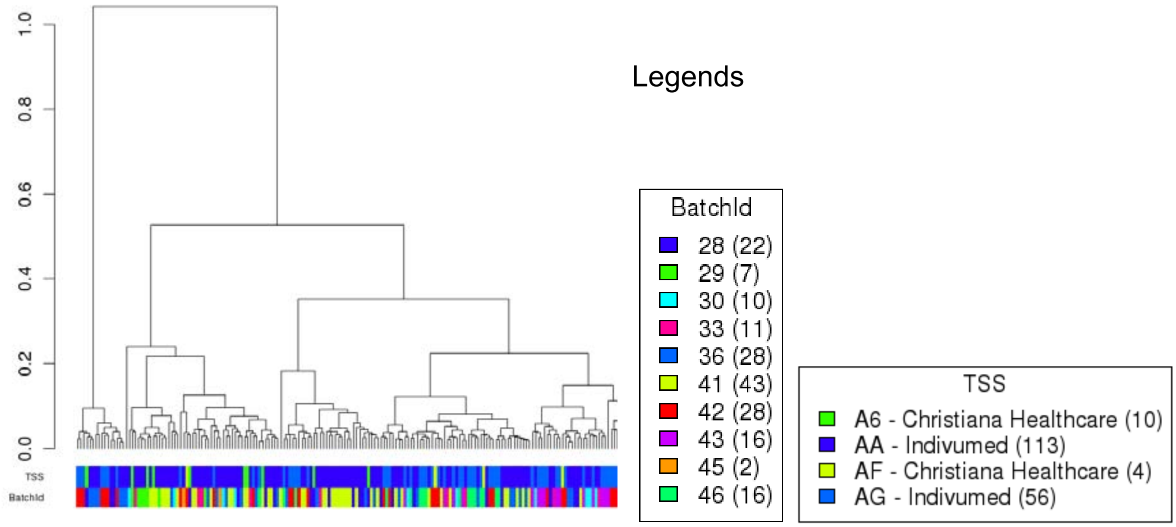




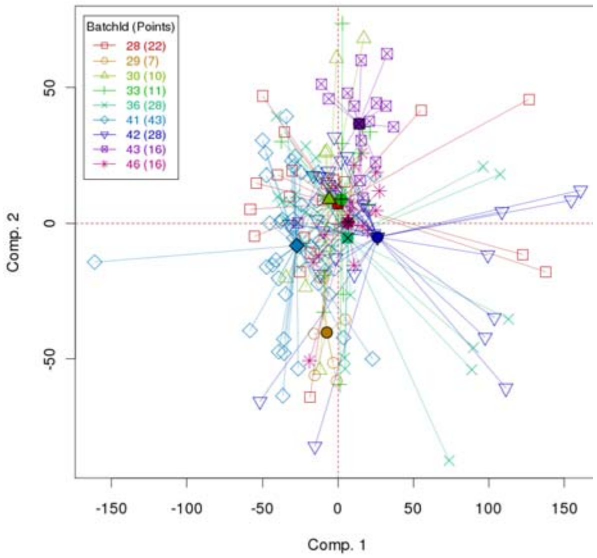
Suppl. Methods Fig. 3. PCA: First two principal components for mRNA expression (microarray), with samples connected by centroids according to batch ID.

Suppl. Methods Fig. 4. PCA: First two principal components for mRNA expression (microarray), with samples connected by centroids according to TSS.

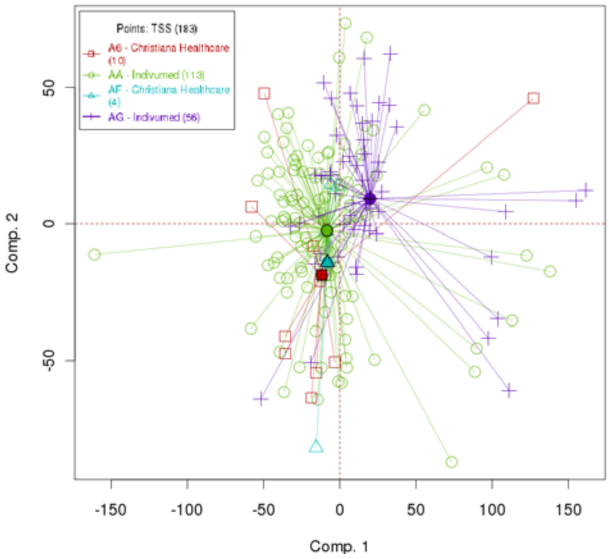
mRNA expression (RNA-seq Illumina GA). The following figures show clustering and PCA plots for the RNA-seq platform. Genes with zero values were removed and the RPKM values were $\log_2$-transformed before generating the figures. Once again, batch 29 (yellow samples on the bottom in Suppl. Methods Fig. 6) stood out from the rest. The other batches or TSSs didn't stand out in either clustering or PCA plots.



Suppl. Methods Fig. 5. Hierarchical clustering for mRNA expression from RNA-seq data



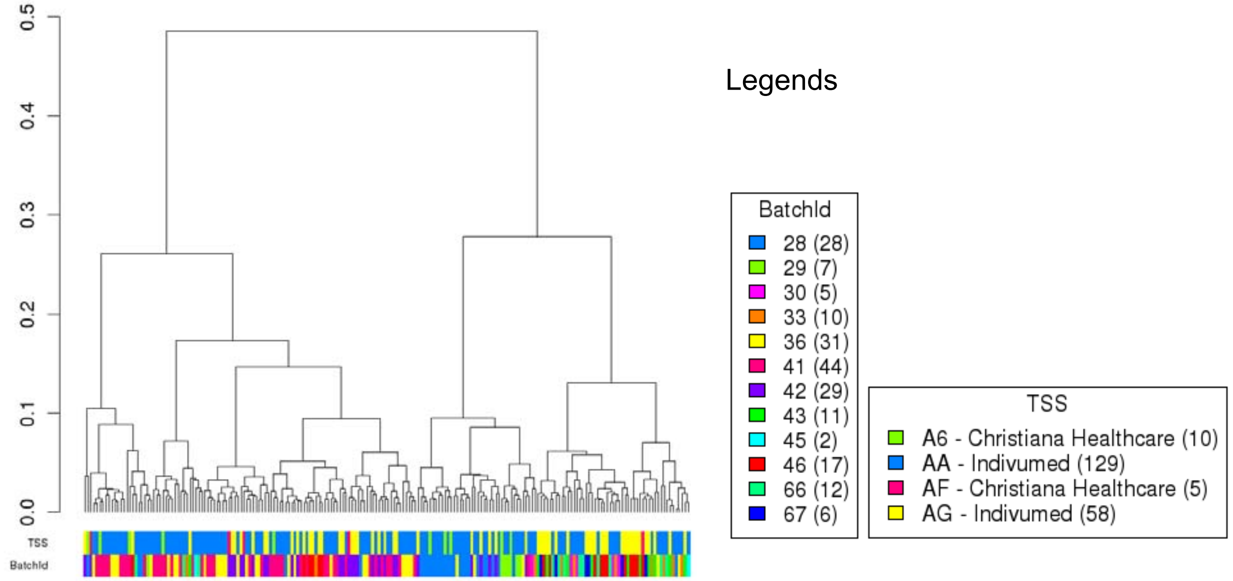
Suppl. Methods Fig. 6. PCA: First two principal components for RNA-seq, with samples connected by centroids according to batch ID.



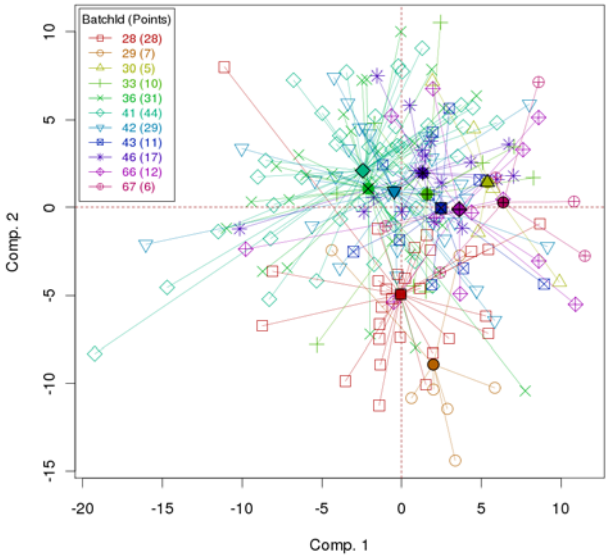
Suppl. Methods Fig. 7. PCA: First two principal components for RNA-seq, with samples connected by centroids according to TSS.

mRNA expression (RNA-seq Illumina GA). The following figures show clustering and PCA plots for RNA-seq miRNA data. Genes with zero values were removed and the read counts were $\log_2$-transformed before generating the figures.

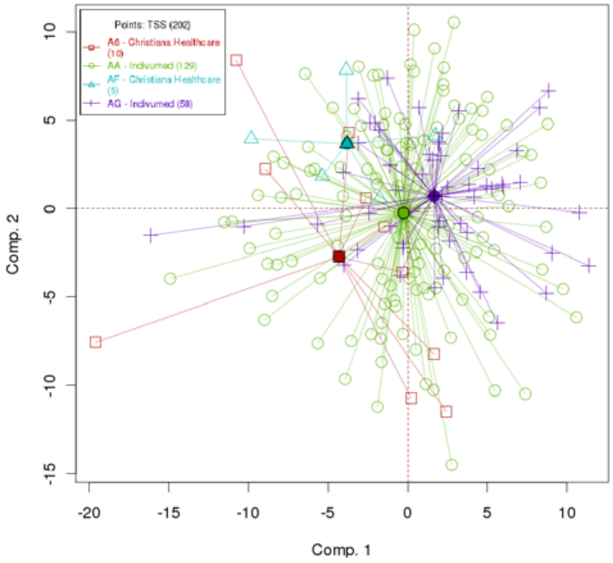
Unlike the other data types, miRNA expression does not show batch 29 to be distinct. MSKCC as a TSS appears distinct in Suppl. Methods Fig. 10 (pink cluster on the left). However, MSKCC does not cluster separately in Suppl. Methods Fig. 8, so we were not too concerned about it. None of the other batches and TSSs stood out in either clustering or PCA plots.



Suppl. Methods Fig. 8. Hierarchical clustering of samples for miRNA expression from RNA-seq data.

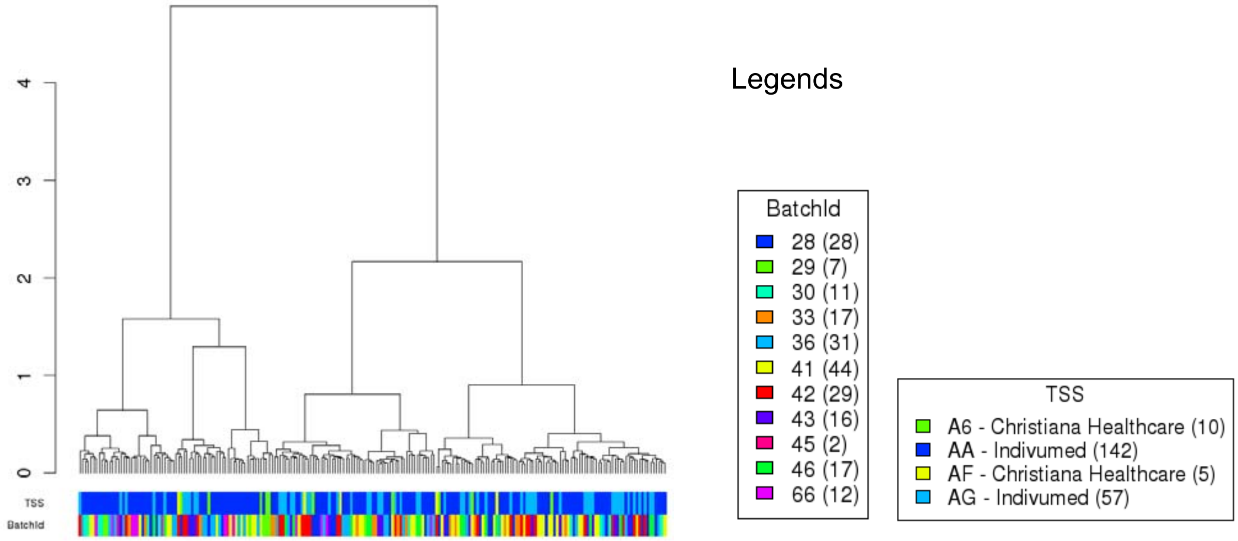


Suppl. Methods Fig. 9. PCA: First two principal components for miRNA expression from RNA-seq data, with samples connected by centroids according to batch ID.

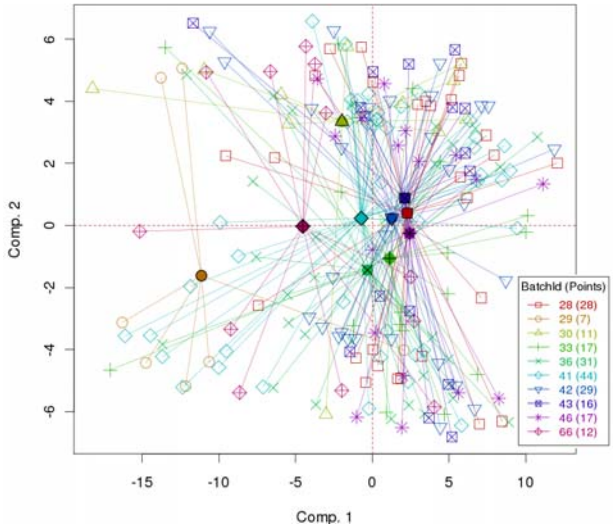


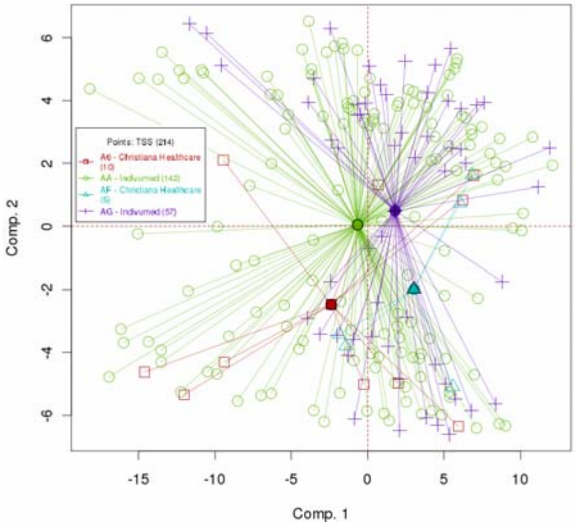
Suppl. Methods Fig. 10. PCA: First two principal components for miRNA expression from RNA-seq data, with samples connected by centroids according to TSS.

DNA Methylation (Infinium HM27 microarray). The following figures show clustering and PCA plots for the Infinium DNA methylation platform. Batch 29 stands out in Suppl. Methods Fig. 12 (but not in Suppl. Methods Fig. 11). None of the other batches or TSSs stood out in either clustering or PCA plots.



Suppl. Methods Fig. 11. Hierarchical clustering plot for DNA methylation data.

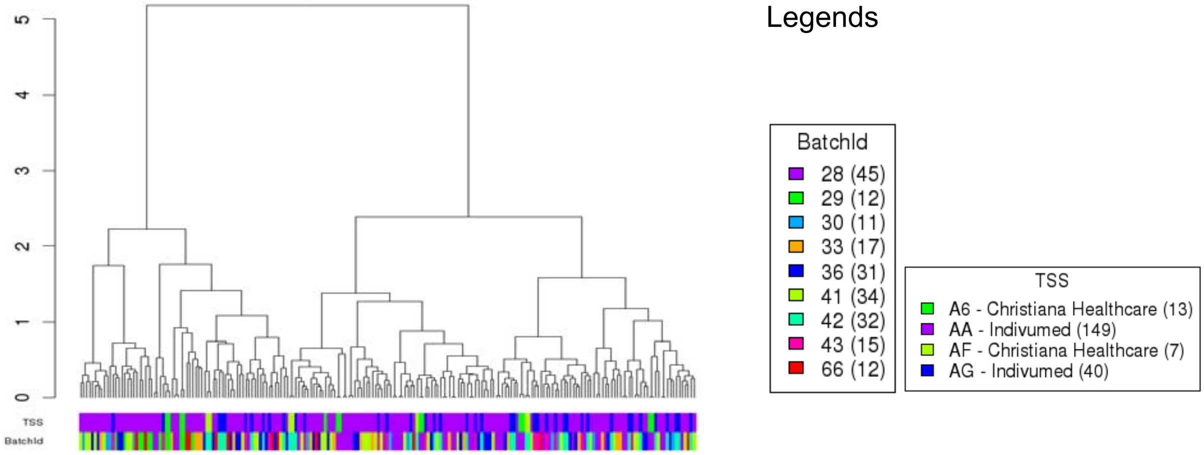




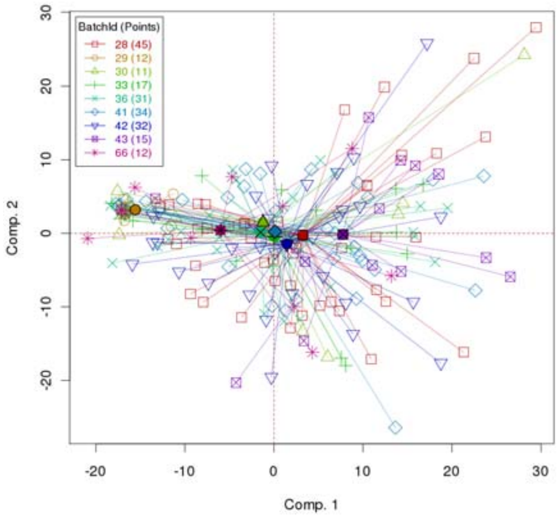
Suppl. Methods Fig. 12. PCA for DNA methylation, with samples connected by centroids according to batch ID.

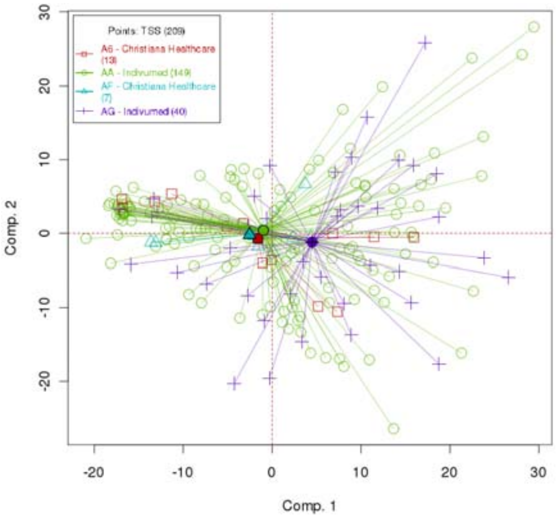
Suppl. Methods Fig. 13. PCA for DNA methylation, with samples connected by centroids according to TSS.

SNPs (GW SNP 6). The following figures show clustering and PCA plots for the SNP platform. At level 3, the TCGA SNP data resemble copy number data when we use chromosomal segment counts (rather than actual SNPs). We mapped the chromosomal segments to genes (using build hg18) and then used them to construct the plots shown in Suppl. Methods Figs. 14-16. Batch 29 once again stands out in the PCA plot in Suppl. Methods Fig. 14 (yellow cluster on the left).



Suppl. Methods Fig. 14. Hierarchical clustering plot for SNP data.





Suppl. Methods Fig. 15. PCA for SNPs, with samples connected by centroids according to batch ID.

Suppl. Methods Fig. 16. PCA for SNPs, with samples connected by centroids according to TSS.

Conclusions. Batch 29 appears distinct from the others in 4 out of 5 data sets: mRNA expression (microarray), mRNA expression (RNASeq), DNA methylation, and SNP data sets. As of the writing of this manuscript, there were 7 samples in batch 29 (except SNP, which had 12). We believe that the differences seen between batch 29 and the others are biological, rather than technical, because (i) batch 29 has been found to consist entirely of MSI/CIMP subtype samples, (ii) batch 29 appears to be distinct across several different platforms, and (iii) both DNA and RNA data types show differences. Consequently, we did not try to apply computational batch effects correction to Batch 29.

None of the other batches or Tissue Source Sites (TSS) in any data set showed consistent batch effects in both clustering and PCA algorithms. Some batches or TSSs stood out in one algorithm or the other, but not both, reducing our concern about them. Based on the above figures, we believe that technical batch effects in the data sets are reasonably small and unlikely to influence high-level analyses in a major way.

Integrated Analysis

Curated Pathway Analysis

We analyzed several pathways that are generally altered in different cancer types, specifically the RAS/PI3K, WNT, TGF-beta, and p53 signaling pathways. For all pathway analyses, we used the set of cases (N=195) with complete data (mRNA expression, DNA copy-number, methylation, and protein mutations). All analyses were further stratified by hypermutation status, resulting in a set of 165 non-hypermutated and 30 hypermutated samples.

We used two approaches to identify altered pathways:

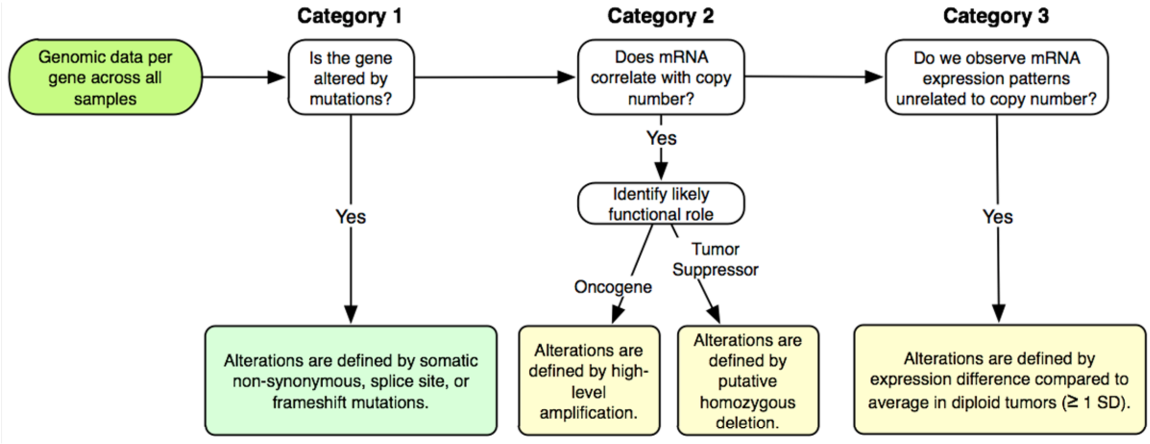
1. An algorithmic approach using MEMo[19]
2. A focused analysis of pathways known to be frequently altered in cancer

Both methods rely on the general abstraction of gene alterations per sample. We used the following approach to determine whether a particular gene was altered or not altered in a particular sample. Our approach was based on first examining each gene across all samples, and binning each gene into one of four categories:

- Category 1: Gene is altered by mutations.
- Category 2: Gene is primarily altered by copy number alterations, and mRNA expression levels correlate with copy number changes.
- Category 3: Gene has evidence of a bimodal expression pattern, unrelated to copy number status.

We then used different alteration criteria for each of the four categories. For example, for Category 2 genes, we classified each gene as a likely oncogene or tumor suppressor, and a gene was called altered in a specific sample if the gene was altered by a high level copy-number amplification or putative homozygous deletion (as defined by GISTIC). Finally, for category 3 genes, alteration status was defined by relative expression compared to the expression distribution in tumor samples diploid in the particular gene, ≥ one standard deviation. In all categories, a gene was called altered if the gene contained a non-synonymous, somatic mutation in a protein-coding region.

A pathway was considered altered in a given sample, if at least one gene in the pathway was altered.


Category 1
Category 2
Category 3
Genomic data per gene across all samples
Is the gene altered by mutations?
Does mRNA correlate with copy number?
Do we observe mRNA expression patterns unrelated to copy number?
Yes
Yes
Yes
Identify likely functional role
Oncogene
Tumor Suppressor
Alterations are defined by somatic non-synonymous, splice site, or frameshift mutations.
Alterations are defined by high-level amplification.
Alterations are defined by putative homozygous deletion.
Alterations are defined by expression difference compared to average in diploid tumors (≥ 1 SD).

Supplementary Methods Figure 17. Assessment of gene alterations used in pathway analysis.

MEMo: Mutual Exclusivity Modules in Cancer. To identify mutually exclusive alterations in colorectal cancer, we ran the MEMo algorithm on the set of non-hypermutated samples. The input data consisted of copy-number altered regions of interest (ROIs) determined by GISTIC, significantly mutated genes determined by MutSig (q < .05), and *IGF2* over-expression status. Because *IGF2* was not present in the default reference network from PathwayCommons (http://www.pathwaycommons.org), we added *IGF2* as a novel node connected to IGF1R. With these settings, MEMo identified only one module with significantly mutually exclusive events (p* < 10-2), which included four genes: *IGF2*, *PIK3CA*, *PTEN*, and *ERBB2* (Supplementary Table 5). This result suggests that *IGF2* could be an activator of the RTK/PI-3-K cascade in colorectal cancer.

Pathway analysis using PARADIGM

Data sets. TCGA COAD and READ data was obtained from the TCGA DCC. TCGA gene expression data was median probe centered within each disease cohort. Within each cohort, data was rank transformed to signed –log10 p-values.

Pathways were obtained in BioPax Level 2 format from http://pid.nci.nih.gov/ and included NCI-PID, Reactome, and BioCarta databases. Interactions from all of these sources were then combined into a merged Superimposed Pathway (SuperPathway). Genes, complexes, and abstract processes (e.g. “cell cycle” and “apoptosis”) were retained and henceforth referred to collectively as pathway concepts. Before merging gene concepts, all gene identifiers were translated into HUGO standard identifiers wherever possible. The belief propagation algorithm employed by PARADIGM can be run with cycles and contradictory interactions. Therefore, for the sake of completeness and simplicity, all interactions were included and no attempt was made to resolve conflicting influences if they existed in the resulting SuperPathway. A breadth-first traversal starting from the concept with the highest number of interactions was performed to build one single component. The resulting merged pathway structure contained a total of 8939 concepts representing 3524 proteins, 4865 complexes, and 550 processes.

Integrated pathway analysis. Integration of copy number, gene expression, and pathway interaction data was performed using the PARADIGM software[20]. Briefly, this procedure infers integrated pathway levels (IPLs) for genes, complexes, and processes using pathway interactions and genomic and functional genomic data from a single patient sample. PARADIGM EM Parameters were trained on the colon and rectal cohorts independently and then combined using sufficient statistic counts.

Clustering genes and patient samples using PARADIGM pathway levels.

To provide an overview of the results of the PARADIGM inference, we generated a clustered heatmap in which we plotted genes by samples according to the inferred pathway activities. First, to avoid any biases in the clustering due to the different sizes of the pathways and differing levels of concepts and abstract processes across the SuperPathway, we extracted the data only for the gene concepts for cluster analysis. The inferences can be interpreted a compact summary of a gene's activity given the copy number, expression and surrounding regulatory neighborhood of the gene.

We first clustered the genes using hierarchical cluster analysis (HCA) as encoded in the Eisen software package[21]. We then produced 50 gene clusters by cutting the HCA tree using the cuttree program available in the R programming environment. Rather than including all of the genes for use in clustering the samples, we instead used mediods from each of the 50 clusters. This procedure avoids any overrepresentation biases that might be present among the genes for driving the sample clustering. For example, large complexes such as the ribosome, or gene families such as the olfactory receptors, can exert a large influence on the sample clustering simply due to the large number of coregulated members. Using the cluster mediods mitigates the effect of such large regulons on sample clustering. We then used the mediods to cluster the tumor samples again using HCA (Figure 5; Supplementary Table 9).

To label the clusters, we performed enrichment analysis on the gene clusters by overlapping the 50 distinct clusters with the sets of genes belonging to the constituent pathways used to build the SuperPathway. We were then able to label each cluster according to the constituent pathway that had the highest representation as determined by a hypergeometric overlap analysis. Genes could then also be labelled by the constituent pathway in which they reside. If a gene resided in multiple constituent pathways then we chose to use the one that was most enriched in the gene's cluster. Samples are also annotated in a similar way based on any clinical information available for the sample. To perform overrepresentation analysis on the clinical information, we forced the samples into ten sample clusters using the cuttree method. Samples could then be annotated with the clinical information that was most significantly enriched in the sample's cluster according to the hypergeometric overlap test (Supplementary Table 9).

To learn what pathway activities might be specific to CRC or shared with other cancers, we compared the CRC PARADIGM results to those obtained for glioblastoma multiform (GBM) and ovarian cysadenocarcinoma (OVCA). GBM and OVCA datasets were obtained and processed in the same manner as the CRC cohort using identical normalization steps and PARADIGM parameter settings. All data for GBM was downloaded on July 2, 2011 and for OVCA on July 3, 2011 from the DCC. This included genome-wide SNP 6 copy number data for 527 GBM and 529 OVCA samples and AgilentG4502A gene expression data for 337 GBM and 439 OVCA tumor samples. 283 GBM and 377 OVCA samples with both expression and copy number were available for integrated analysis at the time of data ingestion. The GBM and OVCA datasets were clustered in the following way. GBM and OVCA clustered datasets were derived using the same procedure as was used for the CRC clusters (i.e. 50 gene cluster mediods were first determined with HCA followed by HCA sample clustering using these mediods). However, the gene clusterings determined by the CRC clustering was then projected onto the GBM and

OVCA datasets so that any coordination of gene activity present in CRC and reflected in GBM or OVCA could be visualized.

Detection of pathway concepts with significantly altered activities
We asked whether a particular pathway concept was found to have inferred activities frequently higher or lower than normal in a significantly high proportion of the cohort of colonic and rectal tumor samples. First, we created a null distribution of pathway concept activities by simulating 1000 random patient samples. The pathway structure was preserved in the simulation so that all topological network properties would be preserved in the random control. Data tuples corresponding to a gene's copy number, expression, methylation, and mutation status were also kept together in the permutation and swapped with whole tuples so that any correlation structure present between the different measurements was preserved in the control. PARADIGM was then run on the randomly generated patient samples to produce a distribution of pathway concept activities. Pathway concepts were excluded from further analysis if they did not obtain a minimum IPL of 0.5 in at least a single patient sample both observed or simulated.

For each sample-concept pair a Z-score (IPZ) was computed between the PARADIGM IPL and the mean and standard deviation of the 1000 IPLs derived from the permuted samples for that particular concept. For each concept, a single modulation score was computed as the mean of the IPZs across the entire cohort for that concept. This yields extreme scores for concepts perturbed in the same direction across the majority of the cohort.

We asked whether these concepts with high modulation scores reside in connected networks. If so it might reflect a common mechanism in colorectal tumors that distinguish the cancer from the normal tissue state. To do this, we retained any regulatory interactions in the SuperPathway that interlinked two concepts found to have a modulation score more extreme than the average absolute level of all modulation scores. This produced a network with 2713 concepts connected by 3687 relations. To visualize the network of modulated pathway concepts, we created a network session that can be loaded into the Cytoscape visualization tool[22]. The node size, color, and opacity were set as functions of each concept's modulation score. All data to reproduce the PARADIGM pathway analysis are available in Supplementary Data File 1 and electronic versions of all Cytoscape sessions are available in Supplementary Data File 2 at http://tcga-data.nci.nih.gov/docs/publications/coadread_2012/.

Pathway signatures for hypermutation and CIMP. One of the stark differences between the samples is the overall mutation frequency. In some samples the rate is as high as thousands per exome while in other samples it is fewer than one hundred. To shed light on the pathways that may either be the result or consequence of the observed hypermutation phenotype, we searched for pathway signatures in the SuperPathway using the TCGA consortium's definitions of hypermuted versus non-hypermutated. One hypothesis is that the CpG island methylation leads to epigenomic silencing of mismatch repair genes, most notably MLH1, which in turn leads to an associated accumulation of mutations. However, a small proportion of patient samples exist that do not fit this general rule. We therefore used the inferred PARADIGM pathway activities to search for mechanisms intrinsic to the hypermutation phenotype irrespective of CIMP status. Vice versa, we also searched for activities associated with CIMP irrespective of hypermutation.

We asked whether we could identify subnetworks of concepts from the SuperPathway that could serve collectively as markers for clinically relevant subgroups of patient samples. Here we

define a “pathway marker map” (PMM) as a subnetwork of interconnected concepts whose activities across the patient samples are correlated, either positively or negatively, with a phenotype of interest. We identified such a map using a two-step process. First, a “marker score” was computed for each concept in the SuperPathway that reflects the degree to which the concept is correlated with CIMP (or hypermutation) status across the cohort of patient samples. Because many variables are correlated with hypermutation and CIMP and obfuscate the correlation of a concept’s inferred activity, we used a linear model to remove the effect of confounding variables.
We applied the following linear model before computing the correlation to hypermutation or CIMP:

$$y_{ij} = \sum_{k=1}^{K} \beta_k x_{ijk} + \varepsilon_{ij} \,,$$

where the β_k coefficients correspond to the intercept and any of K different confounding variables. These included: age at initial pathologic diagnosis, anatomic origin subdivision, distant metastases pathologic spread, gender, histological type, history of colon polyps, loss expression of mismatch repair proteins by IHC, lymph node pathologic spread, number of first degree relatives with cancer diagnosis, primary tumor pathologic spread, prior diagnosis, residual tumor, tumor stage, tumor tissue site, vascular invasion present, the cluster inferred from mRNA transcriptional profiling, the fraction of the genome altered, tumor purity, and finally mutation status in *KRAS*, *BRAF*, *TP53*, *APC*, *PIK3CA*, or *FBXW7*.

A linear model was fit separately for CIMP and hypermutation. For the CIMP model, two additional variables were included as confounders: the hypermutated status as well as an indicator denoting which samples had loss of expression in mismatch repair proteins as determined by immunohistochemistry. For hypermutation markers, the methylation cluster was included as a confounding variable. Thus, the maps derived for CIMP and hypermutation should better reflect mechanisms intrinsic to either CIMP or hypermutation in an effort to tease apart the differences in these phenomena. Marker scores were computed by correlating the residuals derived after fitting the above linear model with either CIMP or hypermutated status.

We then constructed the PMM by collecting all interactions from the SuperPathway that link together a pair of pathway concepts having absolute marker scores greater than average among all concepts. Because the original inferences were derived from an underlying connected network, it is possible that the observed subnetworks of interconnected markers arise from serendipitous but random associations simply due to PARADIGM’s belief propagation framework. We therefore asked whether the observed PMMs are significant. If the interconnection among the genes in the PMM are indeed associated with the CIMP (or hypermutated) status of colorectal tumors, then one would expect that the size of the subnetworks would be larger than those obtained in which the patient data is randomly permuted around the SuperPathway. We indeed find that the observed largest connected sub-pathway is significantly larger than expected by chance for both the CIMP and hypermutation PMMs.

We assessed the significance of the pathway marker maps using randomly simulated patients. We constructed 1000 random assignments of labels to the random patient samples so that the sizes matched those of the observed assignments. We then assessed significance in two ways. First, we asked whether the identified pathway signature had more concepts than expected by chance. Given that a pathway signature was significant in terms of size, we then asked if it interconnected concepts in higher proportions than randomly derived sub-pathways of the same size. The number of incoming and outgoing links for a concept was computed as the concept’s degree. The degree distribution of the observed network was plotted against those randomly generated. We determined if the observed degree distribution was higher than seen in the background by applying a Kolmogorov-Smirnov test to these two distributions.

Data Fusion Method for Identifying Molecular Signatures of Tumor Aggression

Feature Matrix. In preparation for identifying molecular signatures associated with tumor aggressiveness, a “feature matrix” was constructed, with a row for each of the 276 tumor samples and columns containing all available clinical, sample, and molecular data for each sample: protein-coding gene expression levels, microRNA expression levels, copy number alterations, DNA methylation levels, somatic mutations, and PARADIGM integrated pathway levels (IPLs). Data was retrieved from DCC archives dated up to and including Feb 2, 2012. Each column represents a single clinical, sample, or molecular feature. The columns were constructed as follows. Clinical and sample data (78 features): Tumor stage was included as per Supplementary Methods Table1, in which values were derived from AJCC Cancer Staging Manual, 7th edition and TNM classification where available. Tumor stage was further grouped into early stage (Stage I and II) and late stage (Stage III and IV). The MSI values in Supplementary Methods Table 1 were also included. Gene expression (20,503): Gene level RPKM values from RNA-seq (described above, RNA-Seq Data Processing Workflow) were log2 transformed. microRNA expression (705): The summed and normalized microRNA quantification files described above (miRNA-Seq) were log2 transformed. Somatic copy number alterations (106): Broad copy number and focal copy number changes were obtained for peaks identified by GISTIC as described above (SNP Based Copy Number Analysis). DNA methylation (23,094): Probe-specific Level 3 β-values were obtained as described above (DNA Methylation Profiling). Somatic mutations (6,006): From the Mutations Annotation Format (MAF) file, several features were generated for each gene, depending on the type and sequence position of somatic mutations. Mutation types considered were synonymous, missense, nonsense, and frameshift. Protein domains (InterPro) including any of these mutation types were annotated as such, with nonsense and frameshift annotations being propagated to all subsequent protein domains. Each available annotation was used to generate a binary indicator vector indicating whether a particular mutation (e.g. a nonsense or frameshift mutation affecting Armadillo Repeat 1 in APC) is present in a specific sample. Mutation features found in fewer than three tumor samples were removed. PARADIGM (14,168): IPLs (see above: Pathway Analysis using PARADIGM) were used as features. The resulting feature matrix for molecular data types contains 276 rows and 64,582 columns. Thus, the selection for molecular signatures associated with tumor aggressiveness is among 64,582 variables.

For assessing associations between genes and tumor aggressiveness, molecular features were selected on the basis of statistical association with a composite clinical signature consisting of the combination of data for six clinical variables: Lymphatic Invasion Present (No/Yes), Vascular Invasion Present (No/Yes), Histological Type (Non-mucinous or mucinous adenocarcinoma), Fraction Lymphnodes Positive by HE, Tumor Stage, and Distant Metastasis (M0,M1). The composite *p*-value was obtained the using the CDF of the χ^2-distribution with 2·6=12 degrees of freedom, applied weighted version of Fishers' combined statistic for combining *p*-values[23].

$$\chi^2 = -2\sum_{i=1}^{6} w_i \log_e(p_i)$$

Weights w_i were used to balance contributions from each of the clinical variables. For each data type, the reciprocal of the mean value of $\log_e(p_i)$ over all molecular signatures was computed for every clinical variable *i*, and these were then rescaled so as to sum to unity to give the weights

w_i. The individual comparison p-values, p_i, for the association between an individual clinical and molecular feature was computed according to the nature of the data levels of the pair: discrete-discrete (Fisher's exact test); discrete-continuous (ANOVA *F*-test, equivalently *t*-test for binary-continuous) or continuous-continuous (*F*-test). Ranked measurements were used in each case. To assess the possible effect of intra-correlation among clinical variables on the selection procedure we used the truncated product method applied to dependent *p*-values[24]. To transform the *p*-values (Eq.(5), Ref. [24]) the correlation matrix among the values of clinical variables over all samples was used. To account for multiple-testing bias the composite *p*-value was adjusted to minimize false discovery rates using the Benjamini and Hochberg procedure[25]. Features with composite *p*-value less than 0.001 were selected.

A selected variable was classified as a marker for aggressiveness if the clinical data trended towards aggressive signature for every p_i< 0.05, and conversely as non-aggressive marker when opposite trending was observed. For all clinical variables but histological type, aggressiveness corresponds to an increase in the value of that variable. The small minority of molecular signatures (0.8%) that showed an inconsistent trend among clinical variables was discarded.

A software tool to explore the clinical correlates of colorectal cancer in the genomic context is available at http://explorer.cancerregulome.org.

References

1 Umar, A. *et al.* Revised Bethesda Guidelines for hereditary nonpolyposis colorectal cancer (Lynch syndrome) and microsatellite instability. *Journal of the National Cancer Institute* **96**, 261-268 (2004).

2 Stransky, N. *et al.* The mutational landscape of head and neck squamous cell carcinoma. *Science* **333**, 1157-1160, doi:10.1126/science.1208130 (2011).

3 Boland, C. R. & Goel, A. Microsatellite instability in colorectal cancer. *Gastroenterology* **138**, 2073-2087 e2073, doi:10.1053/j.gastro.2009.12.064 (2010).

4 Shinde, D., Lai, Y., Sun, F. & Arnheim, N. Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)n and (A/T)n microsatellites. *Nucleic acids research* **31**, 974-980 (2003).

5 Xi, R. *et al.* Copy number variation detection in whole-genome sequencing data using the Bayesian information criterion. *Proc Natl Acad Sci USA* (In Press 2011).

6 Chen, K. *et al.* BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. *Nature methods* **6**, 677-681 (2009).

7 Beroukhim, R. *et al.* Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 20007-20012 (2007).

8 Mermel, C. H. *et al.* GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome biology* **12**, R41 (2011).

9 TCGA. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).

10 Monti, S., Tamayo, J., Mesirov, J. & Golub, T. R. Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. . *Machine Learning* **52**, 91-118 (2003).

11 Dabney, A. R. ClaNC: point-and-click software for classifying microarrays to nearest centroids. *Bioinformatics (Oxford, England)* **22**, 122-123 (2006).

12 Campan, M., Weisenberger, D. J., Trinh, B. & Laird, P. W. MethyLight. *Methods in molecular biology (Clifton, N.J* **507**, 325-337 (2009).

13 Noushmehr, H. *et al.* Identification of a CpG island methylator phenotype that defines a distinct subgroup of glioma. *Cancer cell* **17**, 510-522 (2010).

14 Houseman, E. A. *et al.* Model-based clustering of DNA methylation array data: a recursive-partitioning algorithm for high-dimensional data arising as a mixture of beta distributions. *BMC bioinformatics* **9**, 365 (2008).

15 Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)* **25**, 1754-1760 (2009).

16 Khattra, J. *et al.* Large-scale production of SAGE libraries from microdissected tissues, flow-sorted cells, and cell lines. *Genome research* **17**, 108-116 (2007).

17 Gaujoux, R. & Seoighe, C. A flexible R package for nonnegative matrix factorization. *BMC bioinformatics* **11**, 367, doi:10.1186/1471-2105-11-367 (2010).

18 Rousseeuw, P. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics* **20**, 56-65 (1986).

19 Ciriello, G., Cerami, E., Sander, C. & Schultz, N. Mutual exclusivity analysis identifies oncogenic network modules. *Genome research*, doi:10.1101/gr.125567.111 (2011).

20 Vaske, C. J. *et al.* Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics (Oxford, England)* **26**, i237-245 (2010).

21 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America* **95**, 14863-14868 (1998).

22 Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome research* **13**, 2498-2504, doi:10.1101/gr.1239303 (2003).

23 Fisher, R. A. Questions and answers #14. *The American Statistician* **2**, 30-31 (1948).

24 Zaykin, D. V., Zhivotovsky, L. A., Westfall, P. H. & Weir, B. S. Truncated product method for combining P-values. *Genetic epidemiology* **22**, 170-185 (2002).

25 Banjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)* **57**, 289-300 (1995).

Supplementary Methods Tables

Supplementary Methods Table 1 - TCGA Platforms

Platform Code	Platform Name	Data Types (Base type-Specific type)
CGH-1x1M_G4447A	Agilent SurePrint G3 Human CGH Microarray Kit 1x1M	CGH-Copy Number Results
HG-CGH-244A	Agilent Human Genome CGH Microarray 244A	CGH-Copy Number Results

HG-CGH-415K_G4124A	Agilent Human Genome CGH Custom Microarray 2x415K	CGH-Copy Number Results
HumanMethylation27	Illumina Infinium Human DNA Methylation 27	DNA Methylation
IlluminaDNAMethylation	Illumina DNA Methylation (OMA002 and OMA003) Cancer Panel I	DNA Methylation
AgilentG4502A_07	Agilent 244K Custom Gene Expression G4502A-07-(1,2, and 3)	Expression-Gene
HT_HG-U133A	Affymetrix HT Human Genome U133 Array Plate Set	Expression-Gene
HuEx-1_0-st-v2	Affymetrix Human Exon 1.0 ST Array	Expression-Gene, Expression-Exon
IlluminaGA_mRNA_DGE	Illumina Genome Analyzer mRNA Digital Gene Expression	Expression-Gene
IlluminaGA_RNASeq	Illumina Genome Analyzer RNA Sequencing	Expression-Gene, Expression-Exon, Expression-Junction
IlluminaGA_miRNASeq	Illumina Genome Analyzer miRNA Sequencing	Expression-miRNA, Expression-miRNA Isoform
H-miRNA_8x15K	Agilent 8 x 15K Human miRNA-specific microarray	Expression-miRNA
IlluminaGA_DNASeq	Illumina Genome Analyzer DNA Sequencing	DNA Sequence Mutations
SOLiD_DNASeq	ABI SOLiD DNA System Sequencing	DNA Sequence Mutations
Genome_Wide_SNP_6	Affymetrix Genome-Wide Human SNP Array 6.0	SNP-Copy Number Results, SNP-SNP, SNP-LOH
Human1MDuo	Illumina Human1M-Duo BeadChip	SNP-Copy Number Results, SNP-SNP, SNP-LOH
HumanHap550	Illumina 550K Infinium HumanHap550 SNP Chip	SNP-Copy Number Results, SNP-SNP, SNP-LOH
bio	Biospecimen Metadata - Complete Set	Clinical-Complete Set
minbio	Biospecimen Metadata - Minimal Set	Clinical-Minimal Set

Definitions of the data type terms in this table may be found online in the TCGA Encyclopedia (https://wiki.nci.nih.gov/x/bCZhAg), or by request to TCGA-DCC-BINF-L@list.nih.gov.

Supplementary Methods Table 2 – TCGA Data Types

Data Type (Base-Specific)	**Level 1 (Raw)**	**Level 2 (Normalized/Processed)**	**Level 3 (Interpreted/Segmented)**	**Level 4 (Summary Finding/Region of Interest results)**
Clinical-Complete Set	Clinical data for 1 patient	NA	NA	NA
Clinical-Minimal Set	Clinical data	NA	NA	NA

	for 1 patient			
CGH-Copy Number Results	Raw signals per probe	Normalized signals for copy number alterations of aggregated regions, per probe or probe set	Copy number alterations for aggregated/segmented regions, per sample	Regions with statistically significant copy number changes across samples
SNP-Copy Number Results	NA	Copy number alterations per probe or probe set	Copy number alterations for aggregated/segmented regions, per sample	Regions with statistically significant copy number changes across samples
SNP-LOH	NA	LOH calls per probe set	Aggregation of regions of LOH per sample	Statistically significant LOH across samples
SNP	Raw signals per probe	Normalized signals *per* probe or probe set and allele calls	NA	Statistically significant SNPs across samples
DNA Methylation	Raw signals per probe	Normalized signals per probe or probe set	Methylated sites/genes per sample	Statistically significant Methylated sites/genes across samples
Expression-Exon	[Array] Raw signals per probe	[Array] Normalized signals per probe or probe set	[Array & RNA-Seq] Expression calls for Exons/Variants per sample	Statistically significant exons/variants across samples
Expression-Gene	[Array] Raw signals *per* probe	[Array] Normalized signals per probe or probe set	[Array & RNA-Seq] Expression calls for Genes per sample	Statistically significant genes across samples
Expression-Junction	NA	NA	Expression calls for splice junctions per sample	Statistically significant splice junctions across samples
Expression-miRNA	[Array] Raw signals per probe	[Array] Normalized signals per probe or probe set	[Array & miRNA-Seq] Expression calls for miRNAs per sample	Statistically significant miRNAs across samples
Expression-miRNA Isoforms	NA	NA	Expression calls for miRNA Isoforms per sample	Statistically significant miRNA Isoforms across samples

DNA Sequence Mutations	NA	Putative mutations	Validated somatic mutations	Statistically significant mutations across samples

Data Types are those listed in Supplementary Methods Tables
Supplementary Methods Table 1. Descriptions of data levels are listed in Supplementary Methods Table 3. Some data levels are not applicable (NA) to particular data types. Italicized data type-data level descriptions indicate that some centers do not produce that data level for its corresponding data type and platform.

Supplementary Methods Table 3 - TCGA Data Levels

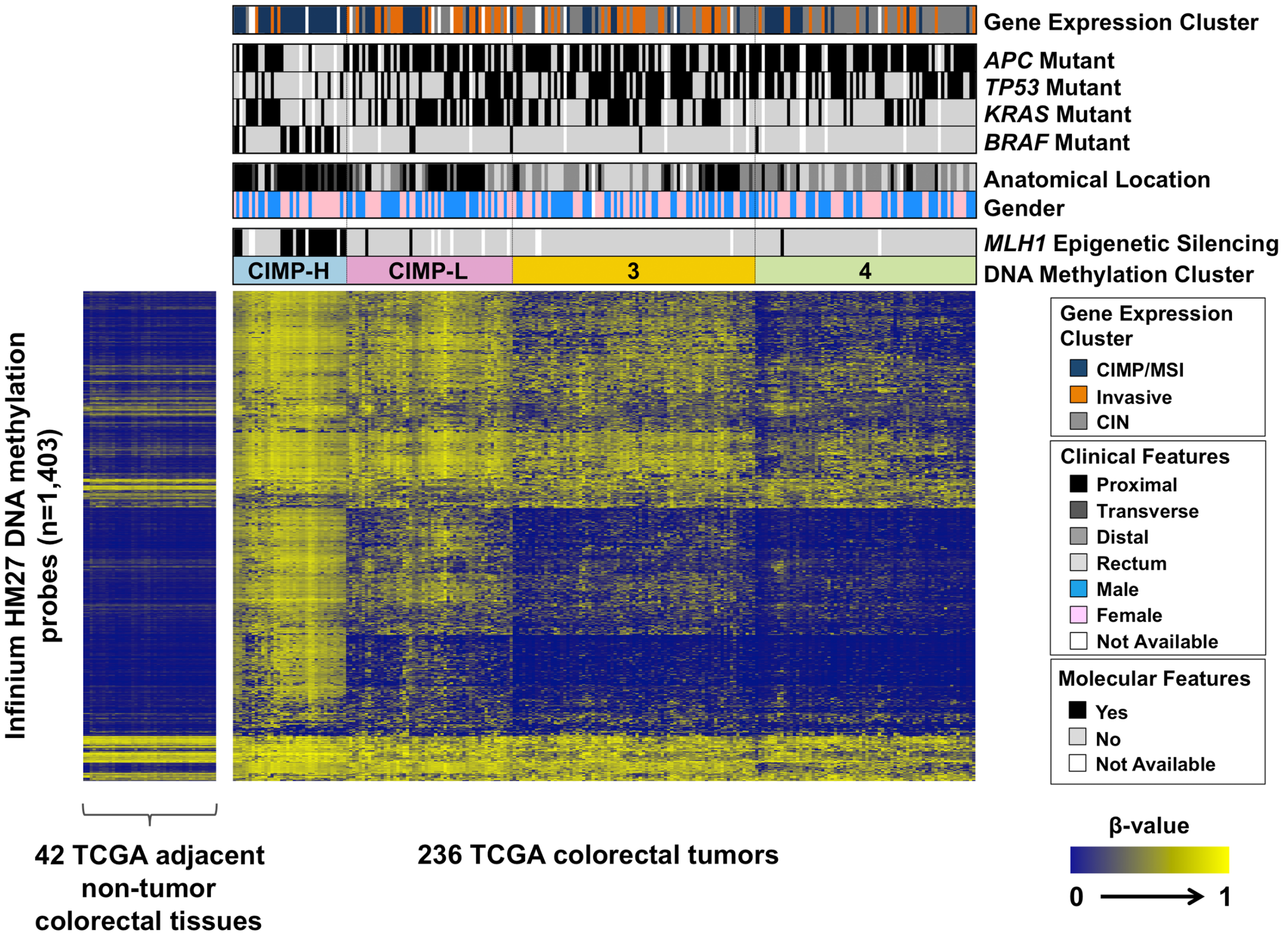
Level Number	Level Type	Description	Example
1	Raw	Low-level data for a single sample, not normalized across samples, and not interpreted for the presence or absence of specific molecular abnormalities.	Affymetrix .CEL file; BAM binary sequence data file
2	Normalized/Processed	Data for a single sample that has been normalized and interpreted for the presence or absence of specific molecular events.	Putative mutation call for a single sample; amplification/deletion/LOH signal for a probed locus in a sample; expression signal of a probe or probe set for a sample
3	Segmented/Interpreted	Data for a single sample that has been further analyzed to aggregate individual probed loci into larger composite or contiguous regions.	Validated mutation call for a single sample; amplification/deletion/LOH signal of a region in the genome for a sample; expression signal of a gene for a sample
4	Summary Finding (ROI)	A quantified association, across classes of samples, among two or more specific molecular abnormalities, sample characteristics, or clinical variables.	A finding that a particular genomic region (a “region of interest”) is found to be amplified in 10% of TCGA glioma samples.

Supplementary Methods Table 4 - miRNA Annotation Priorities

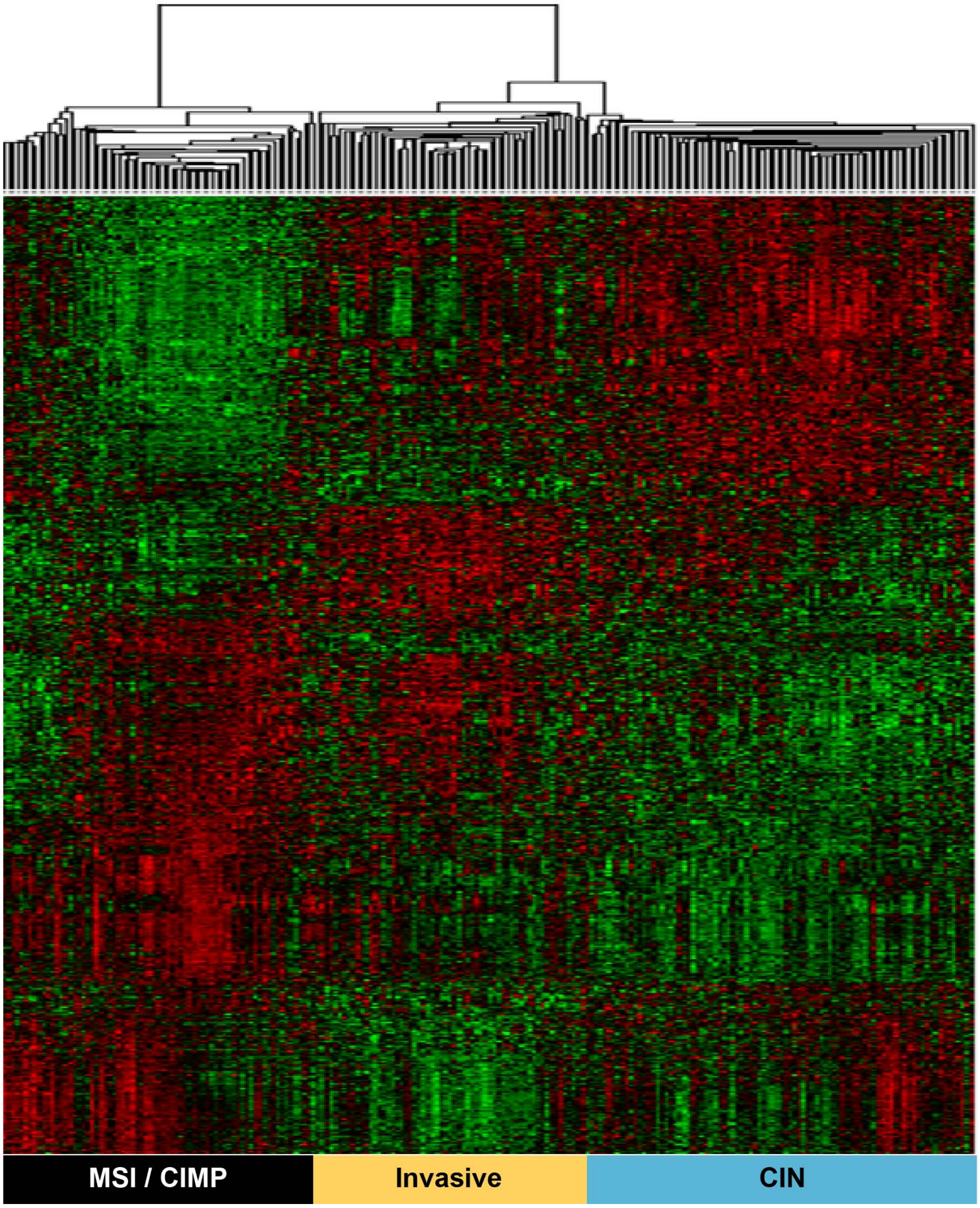
Priority	Annotation type	Database
1	mature strand	miRBase v13
2	star strand	
3	precursor miRNA	
4	stemloop, from 1 to 6 bases outside the mature strand, between the mature and star strands	
5	"unannotated", any region other than the mature strand in miRNAs where there is no star strand annotated	
6	snoRNA	UCSC small RNAs, RepeatMasker
7	tRNA	
8	rRNA	
9	snRNA	
10	scRNA	
11	srpRNA	
12	Other RNA repeats	
13	coding exons with zero annotated CDS region length	UCSC knownGenes
14	3' UTR	
15	5' UTR	
16	coding exon	
17	intron	
18	LINE	UCSC RepeatMasker
19	SINE	
20	LTR	
21	Satellite	
22	RepeatMasker DNA	
23	RepeatMasker Low complexity	
24	RepeatMasker Simple Repeat	
25	RepeatMasker Other	
26	RepeatMasker Unknown	

This table describes the annotation priorities used to resolve multiple database matches of miRNA for a single alignment location and multiple alignment locations for a read.

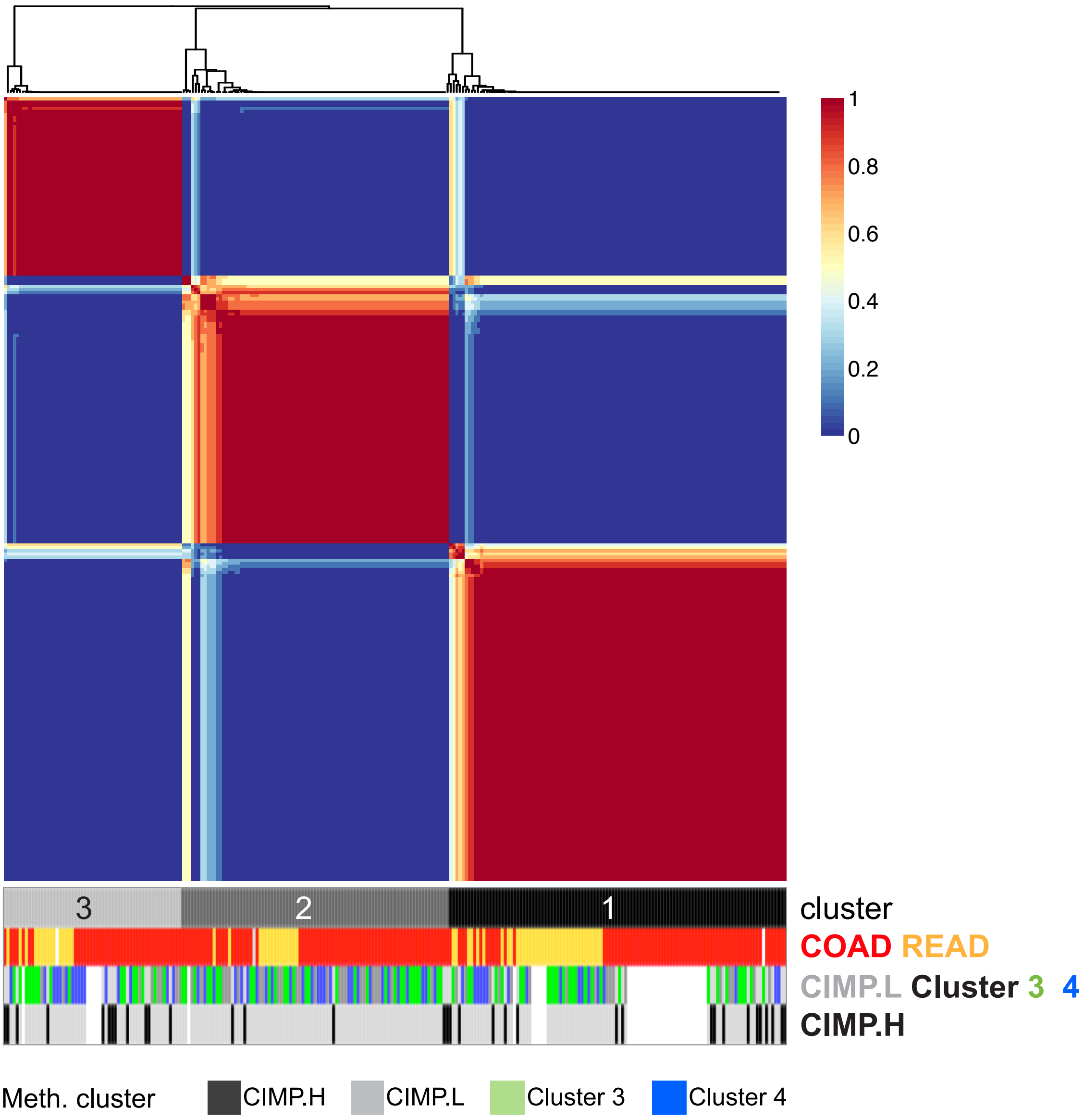
Supplementary Figure 1



Supplementary Figure 2



Supplementary Figure 3



Supplementary figure 4. Sample groups identified from miRNA abundance profiles. Consensus metaheatmap for three sample groups identified from miRNA-seq abundance profiles for 255 COAD/READ tumor samples. NMF consensus clustering (Gaujoux and Seoighe 2010) was applied to a normalized abundance matrix for the 25% most variant mature or star strands (221 MIMATs). The legend colours and dendrogram reflect per-sample cluster membership over 1000 iterations. Tracks under the heatmap show NMF clusters, COAD vs. READ sample types, then sample classifications from DNA methylation clusters: CIMP-H, CIMP-L, cluster 3 and cluster 4.

Method

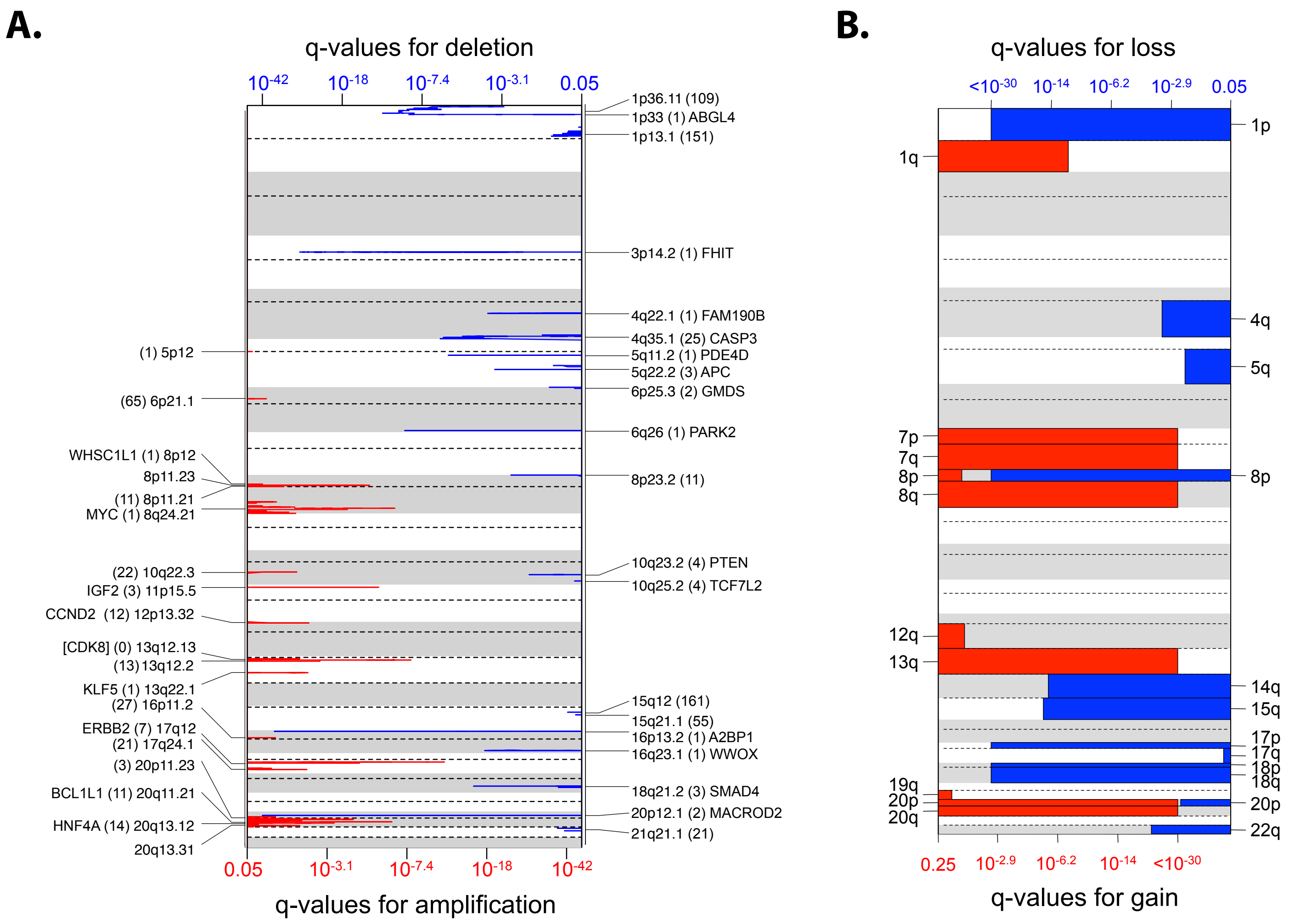
For 2259 TCGA tumor and normal samples, abundance profiles were normalized to 1 M reads mapped to miRBase v16 MIMAT i.e. mature or star strand annotations, and samples were clustered with R v2.12.1's `hclust` function (R Development Core Team, 2011), using a Spearman correlation coefficient as a profile similarity metric. Then, normalized abundance profiles for 255 COAD and READ samples were clustered using 200 iterations of NMF v0.5.2 (Gaujoux and Seoighe 2010), for 3 to 10 clusters, and a silhouette plot (Rousseeuw 1987) was calculated for each cluster result, again using R v2.12.1. Inspection of cophenetic correlation coefficients and overall average silhouette width for each clustering result suggested that 3 clusters was a preferred result, with 5 and 8 clusters also of interest (data not shown). A 3-cluster NMF result was then generated using 1000 iterations, and a silhouette plot was generated for this result, using R 2.13.0.

Gaujoux R, Seoighe C. A flexible R package for nonnegative matrix factorization. BMC Bioinformatics. 2010;11:367.

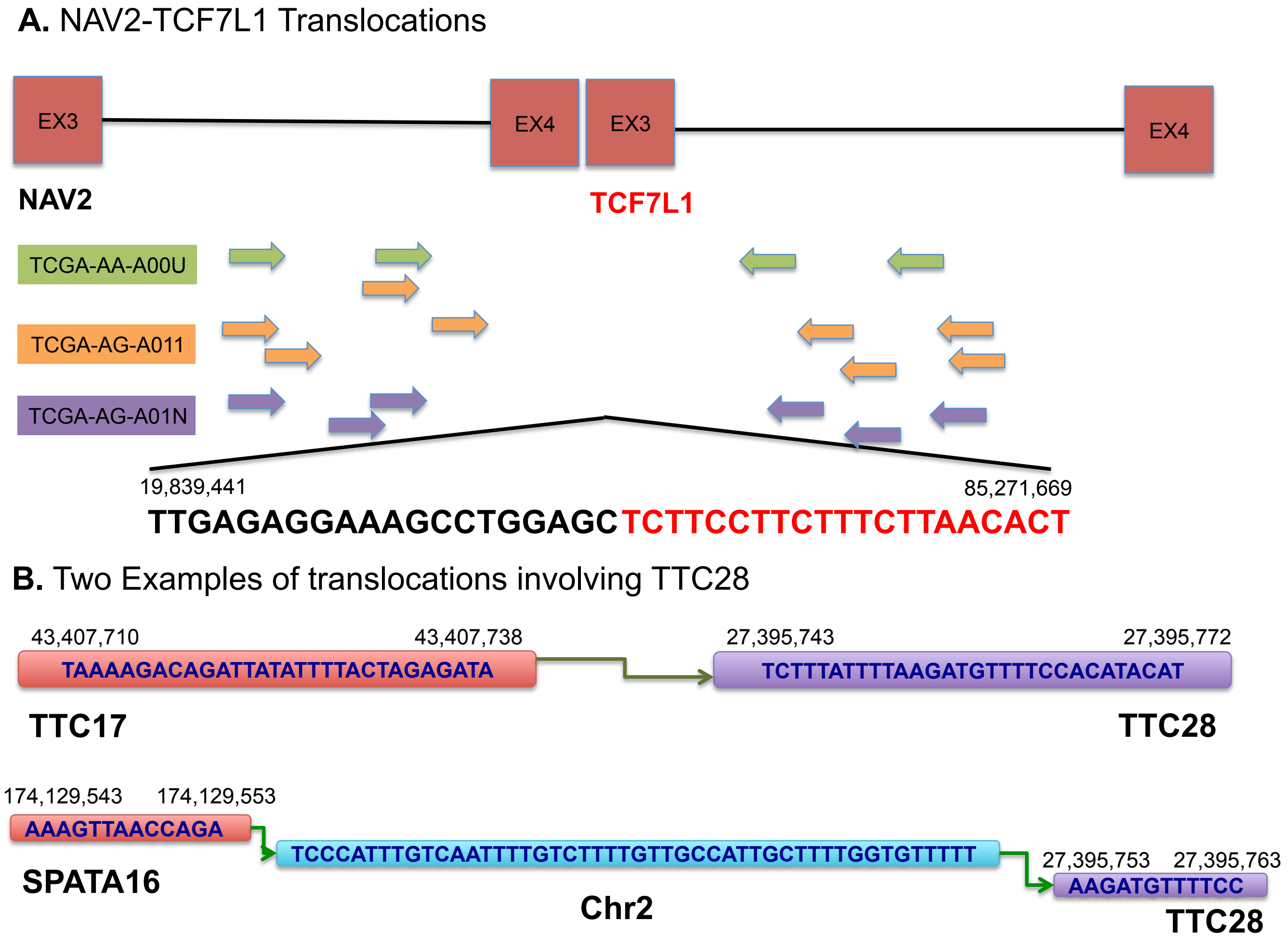
Rousseeuw PJ (1987), "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," Journal of Computational and Applied Mathematics, 20, 53–65.

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, UR L http://www.R-project.org/.

Supplementary Figure 4



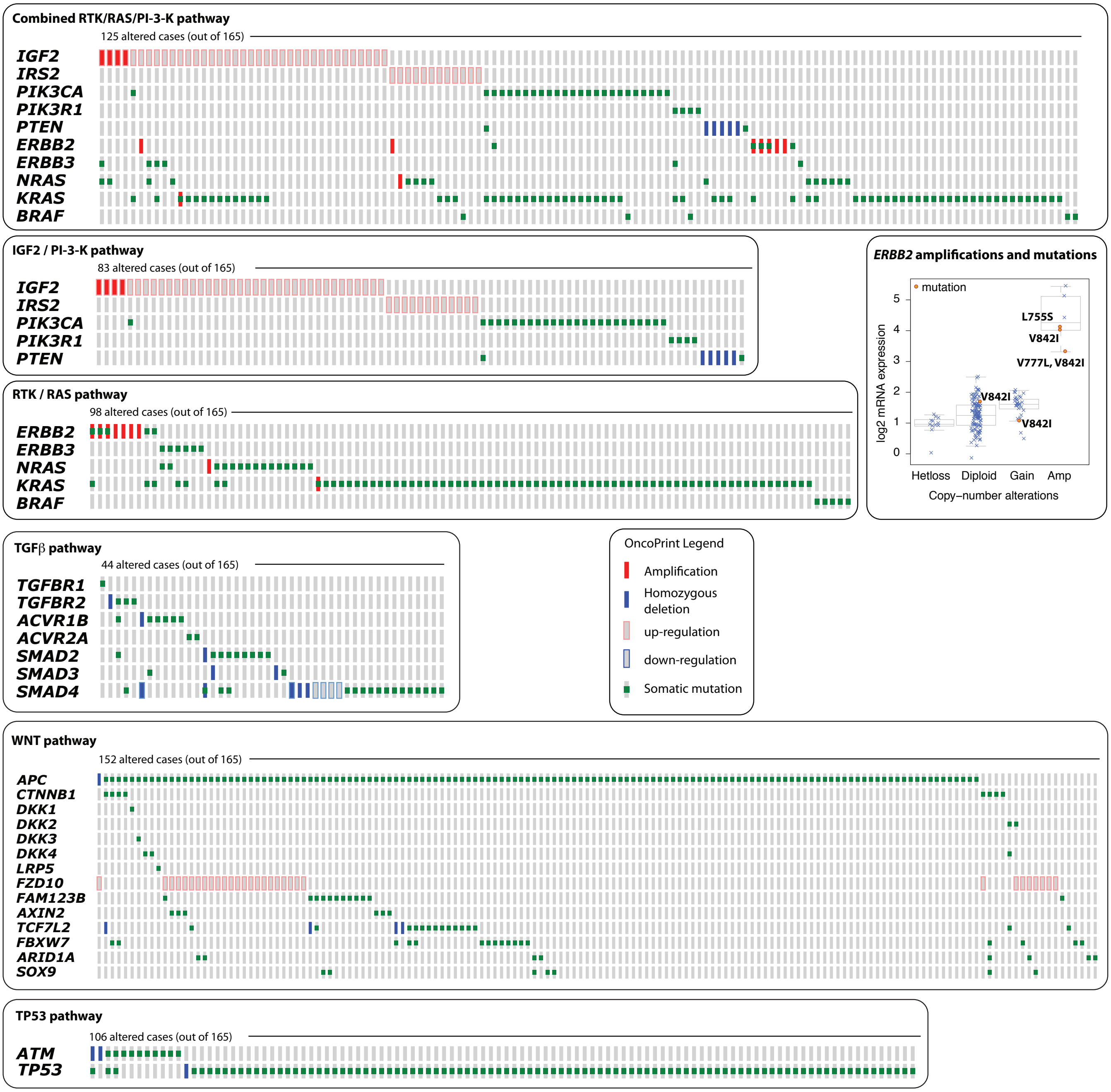
Supplementary Figure 5



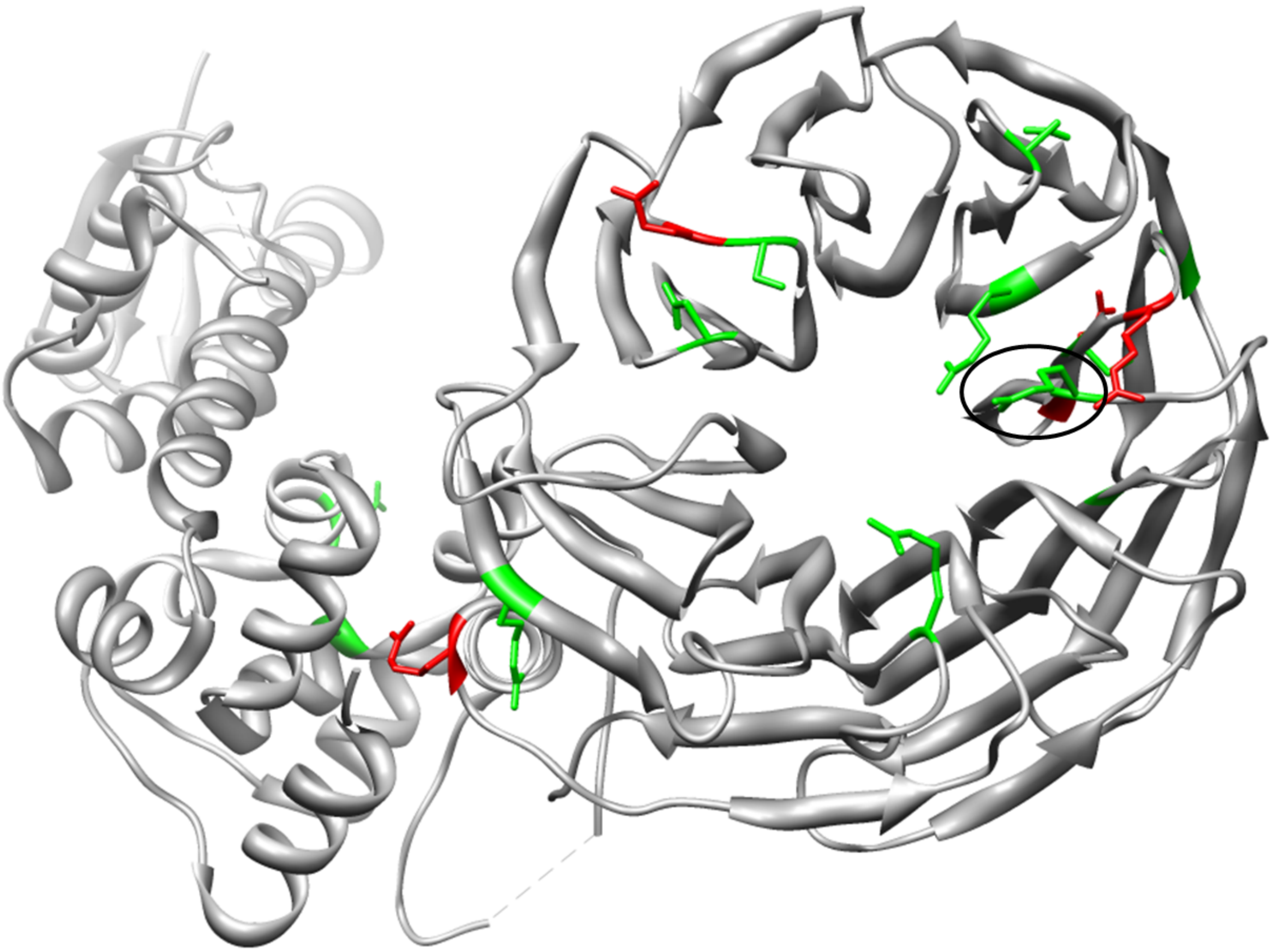
A. NAV2-TCF7L1 Translocations

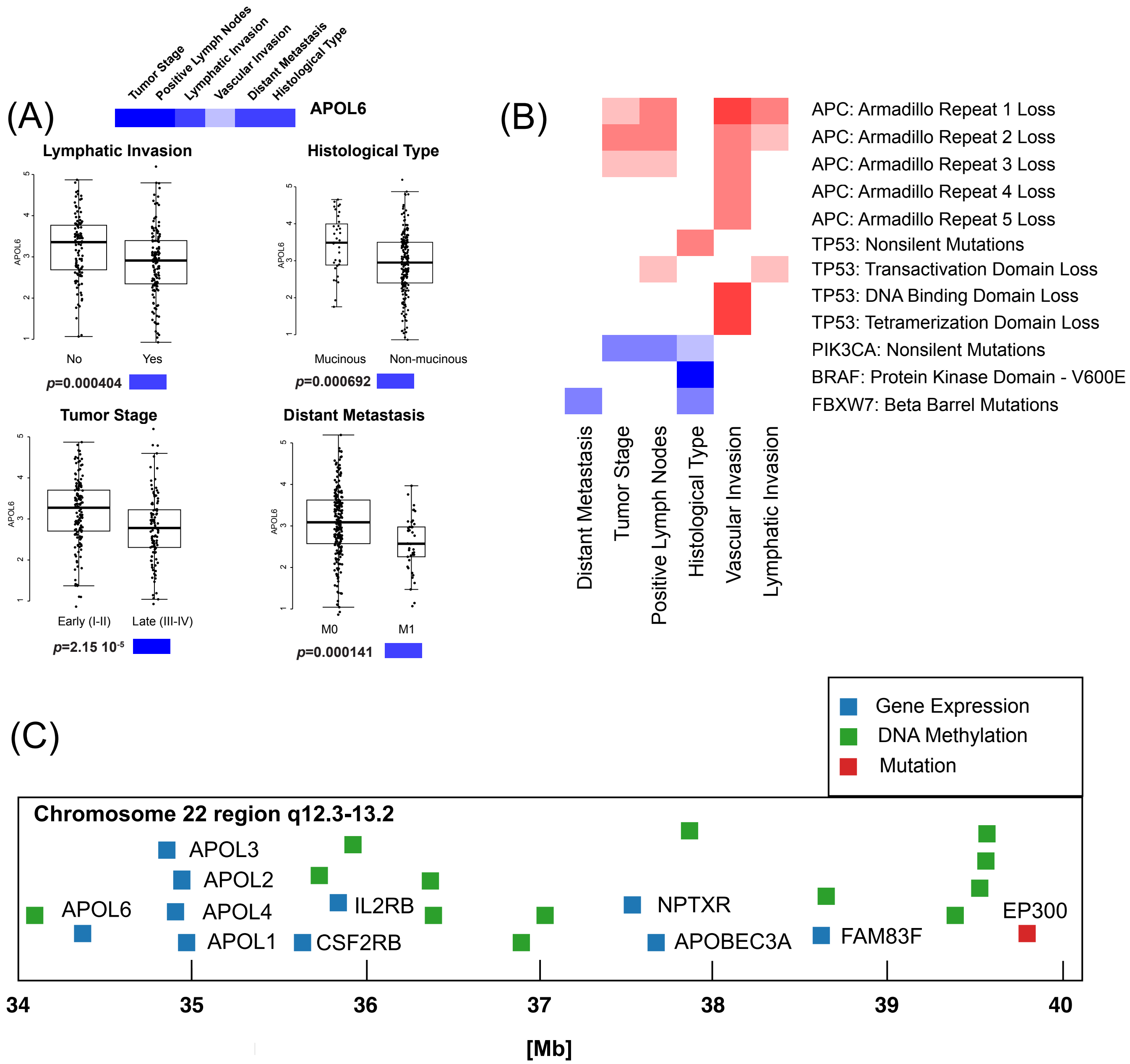
B. Two Examples of translocations involving TTC28

Supplementary Figure 6

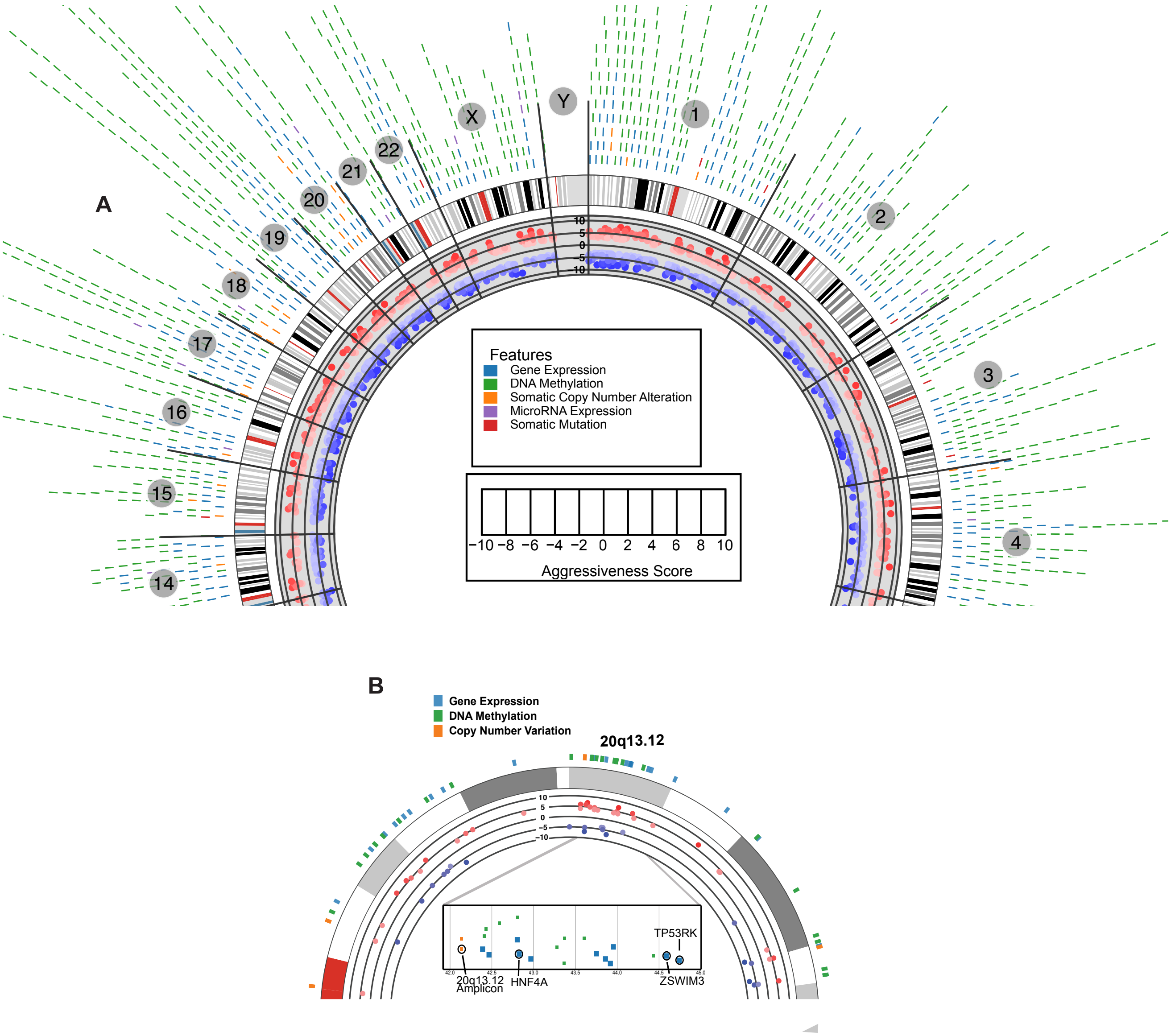


Supplementary Figure 7





Supplementary Figure 8



Supplementary Figure 9