SUPPLEMENTARY METHODS

Exome-capture sequencing

Genomic DNA was extracted using standard methods and pooled libraries of six indexed samples prepared using the TruSeq DNA Sample Preparation Kit (Illumina Inc.). Exome-capture was performed using the TruSeq Exome Enrichment Kit (Illumina Inc.) and 100bp paired-end read sequencing performed on an Illumina HiSeq 2000 System at the Australian Genome Research Facility (AGRF). Raw FASTQ paired-end reads were aligned against the human hg19 reference sequence using the Burrows-Wheeler Alignment (BWA) tool (1). Conversion into bam format was carried out using SAMtools (2). The Picard package was used for sorting and removal of duplicate reads (http://picard.sourceforge.net), and the Genome Analysis Toolkit (GATK) (3) was used for local realignment around regions of nucleotide insertions/deletions (InDels) and recalibration of sequence quality scores. Single nucleotide variants (SNVs) and insertions/deletions were identified using the UnifiedGenotyper application within the GATK package, filtered based on GATK Best Practices quality control filters and annotated against RefSeq using the GATK caller. Variants were annotated against databases of known human germline variations (single nucleotide polymorphism (SNP) database (dbSNP, build 135, SAO = 1), 1000 Genomes Project database (build 20110521), Mills *et al* dataset of small insertions and deletions (4) and germline variants detected in 114 normal colorectal tissues analyzed in our laboratories. Regions of known germline chromosomal segmental duplications were excluded to reduce the

possibility of false-positive variants caused by read mismapping (5). For mutation detection sensitivity/specificity analysis using primary cancers data and the comparison of mutations in cell line pairs, mutations were further filtered to exclude regions with low coverage (defined as <8 reads) in the matched normal or pair. In the comparison of the paired cell lines, when a variant was called in only one member of the pair, the "wild-type" sample was re-inspected for evidence of presence of the variant with a positive call made when the variant was detected in at least two reads.

Transcriptome sequencing

Total RNA was extracted using the Qiagen DNA/RNA isolation kit. cDNA synthesis, library preparation of six indexed samples and RNA-Seq analysis was performed at the AGRF on an Illumina HiSeq2000 to a minimum depth of >100 million paired reads of 50bp or 100bp. Raw reads were assessed for good quality using the FASTQC software. Alignment of transcript sequences to the human reference genome (build hg19) was performed using the TopHat software (6) with default parameters. Gene regions were identified based on alignments of the RefSeq human database by the UCSC genome browser (hg19). Normalized gene expression values were calculated by counting aligning reads per kilobase per million reads mapped (RPKM). Absence of gene expression was defined as a RPKM value of <1. RNA-SEQ data showed good correlation with microarray data for genes with detectable expression on both platforms (**Supplementary Fig. S1**).

SUPPLEMENTARY DATA

Evaluation of mutation detection pipeline

Sensitivity and specificity of our bioinformatics pipeline for identifying somatic mutations in the absence of matched germline data were evaluated by exome-capture sequencing on five non-hypermutated CRCs, five hypermutated MSI-H CRCs and paired normal tissues. Samples were obtained from the Royal Melbourne Hospital via the Victorian Cancer Biobank, Australia. The study was approved by the institutional ethics committee, and all patients gave informed consent. All patients had Dukes stage B adenocarcinoma, and tumor samples were macro-dissected to comprise >80% neoplastic cells.

Results for our custom pipeline using tumor data only were compared to a standard pipeline using paired tumor-normal data. For non-hypermutated and hypermutated MSI-H CRCs, the mean specificity of non-silent mutation detection across cases was 68.8% (range 63.2-75.2%) and 86.5% (range 85.1-88.5%), and the mean sensitivity was 98.4% (range 97.9-100%) and 98.6% (range 97.1-99.9%), respectively. The majority (93.4%) of false-negative calls resulted from somatic mutations mimicking annotated germline variants in reference databases.

Simulation of tumor mutational heterogeneity

Statistical simulations were performed to model the probability distribution of mutational differences between two randomly selected cells from a tumor mass separated by a given number of replications, and compared to the number of mutational differences observed in our paired cell lines. We modelled a basic process of mutations accumulating with no selection (advantage or disadvantage) at a fixed mutation probability in independent tumor cells after all driver mutations have been established. We used mutation probabilities of 10^{-8} per base per cell replication for non-hypermutated and 10^{-6} for hypermutated MSI-H tumors and a target area of 66.6 Mb (representing the genomic region covered by exome-capture sequencing in the diploid human genome) (7). Simulations for both scenarios were performed for a population of 10^5 cells and repeated ten times. Positions of mutations were recorded irrespective of type of mutation. The positions of new mutations were simulated in two stages: the number of mutations each daughter cell acquires was selected from a Poisson distribution with the mean parameter calculated as the number of target bases multiplied by the mutation probability. Then, each new mutation had its position randomly chosen from a uniform distribution over the number of bases; positions not previously mutated for that cell were added to the list of mutated positions for that cell. After each cell replication, half of the cells were randomly chosen to be retained for further simulation in order to prevent an exponential increase in cell number and limit the computational load. Numbers of differences between mutated positions were then tabulated for all possible pairings of a subset of 1000 cells randomly selected from the

population. Simulations were run for 1500 and 110 replications for non-hypermutated and MSI-H hypermutated tumors, respectively. The results from each set of 10 simulations were combined to generate probability distributions for the number of mutational differences with increasing cell replications. These distributions were then used to estimate the minimum number of cell replications required for two randomly chosen cells of a tumor mass to possess the number of mutational differences observed in our paired cell lines at a greater than 99% probability (**Supplementary Table S9**). The code to perform these simulations was written the C programming language. Python scripts were used to extract the numbers of mutation differences at each cell replication and to combine the information from multiple simulation runs.

Simulation of the acquisition of mutations in cell culture

Simulations of the acquisition of mutations in cell culture were performed for the process of serial-passage to estimate the number of cell replications required for any mutation to reach a sequencing detection threshold of 10%. Again, we modelled a basic process of mutations accumulating with no selection in non-hypermutated and hypermutated MSI-H tumors at the fixed mutation probabilities of 10^{-8} and 10^{-6} per base per cell replication, respectively (7). Accumulation of new mutations was simulated for the process of serial passaging, with cell populations replicating from 1×10^5 cells to 2×10^6 cells followed by random selection of 5% of the cells for the next passage. The number of mutations that cells acquire at each replication during passage was modelled as described above. Prior to each simulated split of cells, the proportion of cells harboring any given mutation was tallied and the maximum proportion was recorded.

The entire process was repeated for 400 passages for simulations using the mutation probability of 10^{-8}. For the simulations using the mutation probability of 10^{-6}, all mutations were recorded for the first five passages (corresponding to ~43 replications) after which the target area was restricted to the top 5% most prevalent mutated positions. This restriction was implemented to keep the simulations tractable with respect to computing time and memory resources, and assumes that mutations in new positions are highly unlikely to become the most frequent mutation in the cell population at later passages. A total of 100 passages were simulated for this latter scenario. Each set of simulations was performed five times.

Data were extrapolated to estimate the number of cell replications required for any mutation to reach a sequencing detection threshold of 10%. We observed that after ~10 passages a plot of the logarithm of the proportion of cells with the most frequent mutation against the logarithm of the number of passages yielded an approximately linear trend with the level of variation being approximately constant with passage number (**Fig. 6**). These data were then fitted using linear regression, and 99% prediction intervals constructed for increasing passage numbers.

SUPPLEMENTARY REFERENCES

1. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754-60.

2. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078-9.

3. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297-303.

4. Mills RE, Pittard WS, Mullaney JM, Farooq U, Creasy TH, Mahurkar AA, et al. Natural genetic variation caused by small insertions and deletions in the human genome. Genome Res. 2011;21:830-9.

5. Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, et al. Recent segmental duplications in the human genome. Science. 2002;297:1003-7.

6. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25:1105-11.

7. Tomlinson IP, Novelli MR, Bodmer WF. The mutation rate and cancer. Proc Natl Acad Sci U S A. 1996;93:14800-3.

Supplementary Figures

Supplementary Figure S1. Correlation between RNA-Seq and gene expression microarray data (Affymetrix U133 Plus2.0 Arrays) for six CRC cell lines analyzed in independent experiments. Genes with detectable expression on both platforms were considered, defined as RNA-Seq RPKM value >1 and microarray intensity > log2(100). r-values, Spearman's rank correlation coefficient.

Supplementary Tables

Supplementary Table S1. CRC cell line details and major molecular hallmarks.

Supplementary Table S2. Somatic mutations in protein-coding genes identified in 70 CRC cell lines using exome-capture sequencing.

Supplementary Table S3. Sanger sequencing validation for mutations detected by exome-capture sequencing for 12 genes in 43 CRC cell lines. Genes comprised *APC*, *CTNNB1*, *KRAS*, *BRAF*, *NRAS*, *PIK3CA*, *PTEN*, *SMAD2*, *SMAD3*, *SMAD4, FBXW7* and *TP53*. Cell lines included C32, C70, C80, C84, C99, C106, CACO2, COLO201, COLO205, COLO320-DM, COLO678, DLD1, HCA7, HCC2998, HCT8, HCT15, HCT116, HRA19, HT29, HT55, LIM1215, LIM1899, LIM2099, LIM2405, LIM2551, LOVO, LS174T, LS180, LS411, LS513, RKO, RW2982, RW7213, SW48, SW403, SW480, SW620, SW837, SW948, SW1116, SW1222, T84 and VACO4S.

Supplementary Table S4. Comparison of *POLE* mutation spectra between 9 NSHP and 15 non-NSHP CRC cell lines and TCGA-analyzed primary cancers.

Supplementary Table S5. Significantly altered focal regions of chromosomal deletion or gain/amplification for 63 unique CRC cell lines and 213 TCGA-analyzed primary cancers stratified into non-hypermutated and hypermutated MSI-H cases. Minimal focal regions were deduced using OncoSNP v2.18 suite and GISTIC2.0.

Supplementary Table S6. Levels of mRNA expression for 13 CRC cell lines analyzed using RNA-Seq. Transcript levels are given as reads per kilobase per million mapped (RPKM) values.

Supplementary Table S7. Top 5% of mutated genes based on the proportion of affected samples in CRC cell lines and TCGA-analyzed primary cancers for non-hypermutated and hypermutated MSI-H samples.

Supplementary Table S8. Intersecting top 5% of mutated genes based on the proportion of affected samples between CRC cell lines and TCGA-analyzed primary cancers for non-hypermutated and hypermutated MSI-H samples.

Supplementary Table S9. Simulation of tumor mutational heterogeneity. Estimates are for the minimum number of cell replications required for two randomly chosen cells to possess the number of mutational differences observed in paired cell lines at a greater

than 99% probability. Data were simulated using fixed mutation probabilities of 10^{-8} per base per cell replication for non-hypermutated and 10^{-6} for hypermutated MSI-H tumors.