Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer

Dmitri Mouradov1,2,3, Clare Sloggett4, Robert N. Jorissen1,2,3, Christopher G. Love1,2,3, Shan Li1,2, Antony W. Burgess5, Diego Arango6,7, Robert L. Strausberg8,9, Daniel Buchanan10, Samuel Wormald2,11, Liam O'Connor3,11, Jennifer L. Wilding12, David Bicknell12, Ian P. M. Tomlinson13, Walter F. Bodmer12, John M. Mariadason14 and Oliver M. Sieber1,2,3

1Ludwig Institute for Cancer Research, Parkville, VIC, Australia
2Systems Biology and Personalised Medicine Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
3Faculty of Medicine, Dentistry and Health Sciences, Department of Medical Biology, University of Melbourne, Parkville, VIC, Australia
4VLSCI Life Sciences Computation Centre, a collaboration between Melbourne, Monash and LaTrobe Universities, c/o The University of Melbourne, Carlton, VIC, Australia.
5Structural Biology Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia
6Group of Molecular Oncology, CIBBIM-Nanomedicine, Vall d'Hebron University Hospital, Research Institute (VHIR), Universitat Autònoma de Barcelona, Passeig Vall d'Hebron, 119-129, 08035 Barcelona, Spain
7CIBER de Bioingeniería, Biomateriales y Nanomedicina (CIBER-BBN); Spain;
Ludwig Collaborative Laboratory for Cancer Biology and Therapy, Department of Neurosurgery, Johns Hopkins University School of Medicine, Baltimore, MD

Ludwig Institute for Cancer Research Ltd., New York, NY

Cancer and Population Studies Group, Queensland Institute of Medical Research, Herston, QLD, Australia

Cell and Molecular Biology Laboratory, Systems Biology and Personalised Medicine Division, Walter and Eliza Hall Institute of Medical Research, Parkville, VIC, Australia

Cancer and Immunogenetics Laboratory, Weatherall Institute of Molecular Medicine (WIMM), John Radcliffe Hospital, University of Oxford, Oxford, United Kingdom

Molecular and Population Genetics Laboratory, Wellcome Trust Centre for Human Genetics, Oxford, OX, United Kingdom

Oncogenic Transcription Laboratory, Ludwig Institute for Cancer Research, Austin, VIC, Australia

Corresponding Author: Address requests for reprints to

John Mariadason Oncogenic Transcription Laboratory, Ludwig Institute for Cancer Research Ltd, Austin Hospital, Level 5, Olivia Newton-John Cancer & Wellness Centre, Studley Road, Heidelberg, Victoria, 3084, Australia. Email: John.Mariadason@ludwig.edu.au

or
Oliver Sieber, Colorectal Cancer Genetics Laboratory, Systems Biology and Personalised Medicine Division, Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Victoria, 3052, Australia. E-mail: Oliver.Sieber@wehi.edu.au

RUNNING TITLE: Genomic landscape of colorectal cancer cell lines.

KEYWORDS: colorectal cancer, cell lines, exome sequencing, mutation, gene expression

WORD COUNT: 4,179

FIGURES: 6

TABLES: 0

COMPETING INTERESTS: None.

FINANCIAL SUPPORT: Supported by Cancer Australia through a Project Grant (APP1030098; to OMS), Ludwig Institute for Cancer Research through a Project Grant (to JMM, AAC, OMS), NHMRC Overseas Postdoctoral Fellowship (519795; to SW) and a Victorian State Government Operational Infrastructure Support grant. JMM holds an Australian Research Council Future Fellowship.
ABSTRACT

Human colorectal cancer (CRC) cell lines are used widely to investigate tumor biology, experimental therapy and biomarkers. However, to what extent these established cell lines represent and maintain the genetic diversity of primary cancers is uncertain. In this study, we profiled 70 CRC cell lines for mutations and DNA copy-number by whole-exome sequencing and SNP microarray analyses, respectively. Gene expression was defined using RNA-Seq. Cell line data was compared to that published for primary CRCs in the Cancer Genome Atlas. Notably, we found that exome mutation and DNA copy-number spectra in CRC cell lines closely resembled those seen in primary colorectal tumors. Similarities included the presence of two hypermutation phenotypes, as defined by signatures for defective DNA mismatch repair and DNA polymerase ε (POLE) proofreading deficiency, along with concordant mutation profiles in the broadly altered WNT, MAPK, PI3K, TGFβ and p53 pathways. Further, we documented mutations enriched in genes involved in chromatin remodeling (ARID1A, CHD6, SRCAP) and histone methylation or acetylation (ASH1L, EP300, EP400, MLL2, MLL3, PRDM2, TRRAP). Chromosomal instability was prevalent in non-hypermutated cases, with similar patterns of chromosomal gains and losses. While paired cell lines derived from the same tumor exhibited considerable mutation and DNA copy-number differences, in silico simulations suggest that these differences mainly reflected a pre-existing heterogeneity in the tumor cells. In conclusion, our results establish that human CRC lines are representative of the main subtypes of primary tumors at the genomic level, further validating their utility as tools to investigate CRC biology and drug responses.
INTRODUCTION

Colorectal cancer (CRC) is a leading cause of cancer-related morbidity and mortality (1). Human CRC cell lines are an important, commonly-used preclinical model system for studying this disease, and have provided essential insights into tumor molecular and cell biology. Cell lines are a fundamental tool used in the discovery of new anti-tumor compounds and for the discovery of drug sensitivity, resistance and toxicity biomarkers, with molecular markers of response to conventional chemotherapies and targeted agents showing clinical utility in patients (2-7). Examples include the relationship between tumor microsatellite instability-high (MSI-H) status and lack of 5-fluorouracil (5FU) response (2), and mutations in KRAS, BRAF and PIK3CA exon 20 and resistance to anti-EGFR antibody therapy (3). However, to what extent CRC cell lines represent and maintain the genetic diversity of primary cancers remains controversial.

Over the past two decades, major cancer genes and pathways central to CRC development have been delineated, including the WNT, MAPK, PI3K, TGF-β and p53 pathways. Two broad molecular subtypes of CRC have emerged, characterized by MSI-H (~15%) or chromosomal instability (CIN, ~60%) (8-10). MSI-H CRCs occur predominantly in the proximal colon, and often show poor differentiation, mucinous histology and increased peritumoral lymphocytic infiltration (11). These tumors exhibit hypermutation caused by defective DNA mismatch repair (MMR), tend to be near-diploid and to have a CpG island methylator phenotype (CIMP), and harbor mutations in a distinct set of driver genes including BRAF, PTEN and TGFB1 (12-14). In contrast, chromosomally unstable tumors are more common in the distal colorectum and tend to
develop along the classical genetic pathway of colorectal tumorigenesis, with mutations in *APC*, *KRAS*, *SMAD4* and *TP53* (15).

Recent cancer genome sequencing studies have revealed additional details of the genetic profiles of human CRC, highlighting their diversity. An initial whole-exome sequencing study on 11 microsatellite stable (MSS) CRCs demonstrated that such tumors harbor ~80 coding sequence mutations with a small number of commonly mutated driver genes and a large number of “private” mutations (16). Subsequently, The Cancer Genome Atlas (TCGA) Network reported comprehensive data on 223 unselected sporadic CRCs (17). Hypermutation was identified in ~15% of carcinomas, with three-quarters of these displaying the expected MSI-high (MSI-H) phenotype associated with epigenetic silencing and/or mutation of MMR genes. However, one-quarter of hypermutated tumors did not show MSI-H and appeared to be associated with DNA polymerase ε (POLE) mutations, perhaps representing a distinct CRC subtype. Twenty-four genes were identified as significantly mutated in CRC including several novel candidate genes such as *ATM*, *ARID1A*, *TCF7L2*, *SOX9* and *FAM123B*. A number of recurrent DNA copy-number alterations were reported comprising potentially drug-targetable amplifications of *ERBB2* and *IGF2*. A similar study on 74 pairs of primary human colon tumors reported highly concordant results, and in addition identified low frequency fusion transcripts involving R-spondin family members *RSPO2* and *RSPO3* (18).

Here, we report the first comprehensive whole-exome and DNA-copy number analyses of 70 of the most widely used colorectal cancer (CRC) cell lines, and undertake a detailed comparison of genetic alterations with those reported for TCGA-
analyzed primary cancers. We demonstrate that the spectra of mutations and DNA copy-number aberrations in CRC cell lines are representative of primary tumors, including hypermutation phenotypes and targeting of major signaling pathways. Our data further highlight novel aspects of CRC biology, including significant enrichment of mutated genes involved in chromatin remodeling and histone methylation or acetylation. Our results verify established CRC cell lines as a useful preclinical model system, and provide a comprehensive genomic data resource enabling informed choices when selecting cell lines for functional and pharmacogenomics research.

MATERIALS AND METHODS

Colorectal cancer cell lines and TCGA-analyzed primary cancers

The 70 CRC cell lines studied were obtained from a range of sources, listed below, over a period spanning several years (Supplementary Table S1). C10, C106, C125, C135, C32, C70, C80, C84 and C99 were generated by the W.F. Bodmer laboratory, CACO2, COLO201, COLO205, COLO320-DM, DLD1, HCC2998, HCT116, HCT15, HCT8, HT29, LOVO, LS174T, LS180, LS411, LS513, NCI-H716, NCI-H747, RKO, SKCO-1, SNU-C1, SNU-C2B, SW1116, SW1417, SW1463, SW403, SW48, SW480, SW620, SW837, SW948 and T84 were obtained from the ATCC, KM12 was obtained from DCTD, COLO678 and CX-1 were obtained from DSMZ, Gp2D, Gp5D, HT115 and HT55 were obtained from ECACC, CCK81 and CoCM-1 were obtained from HSRRB, VACO10 and VACO4S were provided by Dr John A. McBain (University of
Wisconsin School of Medicine, USA), SNU-175, SNU-283 and SNU-C4 were obtained from KCLB, LIM1215, LIM1899, LIM2099, LIM2405 and LIM2551 were obtained from the Ludwig Institute for Cancer Research, SW1222 were provided by Prof Meenhard Herlyn (The Wistar Institute, USA), HDC101, HDC54, HDC82, HDC87 and HDC90 were provided by Prof. Martin Schwab (DKFZ, Germany), HCA46, HCA7 and HRA19 were provided by Dr. S. C. Kirkland (Royal Postgraduate Medical School, United Kingdom), and RW2982 and RW7213 were provided by Dr Paul Calabresi (Roger Williams General Hospital, USA). Cells were cultured with Dulbecco's modified Eagle's medium and 10% FBS at 37°C and 5% CO2. Cell line authentication was performed using the Promega StemElite ID System at the Queensland Institute for Medical Research (QMIR) DNA Sequencing and Fragment Analysis Facility (January 2013). Cells were tested for mycoplasma by the MycoAlert™ Mycoplasma Detection Kit (Lonza). Published exome-capture sequencing data on 223 CRC patients were retrieved from the TCGA (19). SNP array data were available for 213 of these patients. Mutation data were filtered for exons and at splice-sites (+/-3bp).

**Exome-capture sequencing**

Exome-capture was performed using the TruSeq Exome Enrichment Kit (Illumina) and 100bp paired-end read sequencing performed on an Illumina HiSeq 2000 at the Australian Genome Research Facility (AGRF). Variant detection was implemented using a modified GATK Best Practice protocol; variants were filtered against known germline variations and those detected in 114 in-house normal colorectal
tissues (Supplementary Methods). Regions of known germline chromosomal segmental duplications were excluded (20).

**Transcriptome sequencing**

RNA-Seq analysis was performed at the AGRF on an Illumina HiSeq2000 to a depth of >100 million paired reads. Alignment to the human reference genome (hg19) was achieved using TopHat (21), and RPKM values calculated. Absence of gene expression was defined as a RPKM value of <1 (Supplementary Methods).

**SNP microarray analysis**

SNP assays were performed by the AGRF using Illumina Human610-Quad BeadChips (GSE55832) and processed using GenomeStudio (Illumina). SNPs showing germline alterations, based on 637 normal samples, were excluded. DNA copy-number segmentation was performed using OncoSNPv2.18 (22). Regions of significantly altered genome were identified using GISTIC2.0 (23).

**Microsatellite instability analysis**

MSI analysis was performed for the National Cancer Institute-recommended microsatellite marker panel BAT25, BAT26, D2S123, D5S346 and D17S250 using fluorescently-labelled primers on a 3130xl Genetic Analyzer (Applied Biosystems) (24).
MSI-H was diagnosed if instability was evident at two or more markers.

**CIMP marker and MLH1 promoter methylation analysis**

MethyLight real-time PCR was performed for the Weisenberger *et al* CIMP marker panel (IGF2, SOCS1, RUNX3, CACNA1G, NGN1), *MLH1*, and the reference gene ALU (C-4) (25). The percentage of methylated reference (PMR) was calculated for GENE:ALU ratio of template amount in a sample against GENE:ALU ratio of template amount in methylated reference DNA. Samples with a PMR greater than 10 for ≥3 CIMP markers were classified as CIMP-positive.

**Statistical Analysis**

Statistical analyses were conducted using the R statistical computing software (http://www.R-project.org/). Differences between groups were assessed using Fisher's exact test or chi-squared test for categorical variables and the Student's t-test for continuous variables. Correlation between RNA-Seq and microarray gene expression data were assessed using Spearman's rank correlation coefficient. Overlap of gene lists was assessed using a hypergeometric test. GISTIC2.0 analysis for significantly altered focal regions of chromosomal deletion or gain/amplification were adjusted for multiple testing, implementing a BH-FDR adjusted Q-value cut-off of 0.05. All statistical analyses were two-sided and considered significant when P<0.05.
RESULTS

Exome mutation profiles in 70 CRC cell lines

Seventy commonly used CRC cell lines, comprising 63 unique lines and 7 duplicate lines, were analyzed for mutations in protein-coding exons and exon-intron boundaries by whole-exome capture sequencing (Supplementary Table S2). In all cases, >20-fold coverage of >80% of targeted exons was achieved. As matched normal tissue for these cell lines was not available, putative somatic mutations were identified by annotation against databases of known human germline variants. Regions of known germline chromosomal segmental duplications were excluded to reduce the possibility of false-positive variants caused by read mismapping. Mean pipeline sensitivity and specificity for non-silent variants were shown to be 79.2% and 98.6% in an analysis of 10 primary CRCs with and without paired normal tissue, with the majority (93.4%) of false-negative calls resulting from somatic mutations mimicking annotated germline variants (Supplementary Data). Accuracy of mutation calling was verified by Sanger re-sequencing for twelve selected genes (APC, CTNNB1, KRAS, BRAF, NRAS, PIK3CA, PTEN, SMAD2, SMAD3, SMAD4, FBXW7 and TP53) on 43 cell lines, with validation of 97.6% (165/169) of mutations (Supplementary Table S3).

For the 63 unique cell lines, the total number of putative somatic mutations ranged from 219 to 8657 (mean=1498.3). Similar to the primary CRCs reported by The TCGA (17), the prevalence of mutations varied substantially, ranging from 6.6 to 260.0
per 10^6 bases, with evidence for non-hypermutated and hypermutated groups (Fig. 1). Cell lines with a mutation prevalence of >25 per 10^6 bases were designated as hypermutated. Hypermutation was observed in 34.9% (22/63) of cell lines as compared to 15.7% (35/223) of the TCGA cancers (P<0.001). 86.3% (19/22) of hypermutated cell lines exhibited MSI-H, similar to the proportion found in cancers (80.0%, 28/35, P=0.725). MSI-H status in cell lines was associated with MLH1 hypermethylation, mutations in MMR genes and presence of CIMP (P<0.009, Fig. 1). Consistent with the established association between MSI-H and proximal tumor location in CRC, all MSI-H cell lines with available site details originated from the proximal colon (Supplementary Table S1). There was no evidence of imbalance in site distribution between cell lines and cancers (P=0.937).

**Mutation spectra in hypermutated CRC cell lines**

Compared to non-hypermutated cell lines, MSI-H cell lines displayed a higher number of InDels (41.1-fold; P<0.001) and SNVs (9.7-fold; P<0.001) as observed in the TCGA cancers (InDels: 18.3-fold, SNVs: 10.8-fold; P<0.001) (Fig. 2A-B). As expected, InDels in MSI-H cases tended to occur at nucleotide-repeat (≥N_5 or ≥NN_3) sequences (MSI-H vs. non-hypermutated: cell lines 82.7% vs. 17.9%, cancers 73.4% vs. 38.4%; P<0.001), a bias not observed for the SNVs (cell lines 2.9% vs. 3.7%, cancers 2.3% vs. 2.9%) (Supplementary Table S2). The SNV spectrum differed significantly between MSI-H and non-hypermutated cases, with an increased proportion of A>G/T>C transitions (cell lines 1.5-fold, cancers 1.8-fold; P<0.011) and decreased C>G/G>C
transversions (cell lines -4.5-fold, cancers -5.2-fold; P<0.001) (Fig. 2C).

A second type of hypermutated tumor was observed among cell lines and the TCGA cancers. These cases were MSS and compared to non-hypermutated cases exhibited a nucleotide-substitution hypermutator phenotype (NSHP) characterized by a substantial increase of SNVs (cell lines 16.8-fold, cancers 56.8-fold; P<0.005) (Fig. 2A-B). The SNV spectrum in 2/3 NSHP cell lines (HT115, HCC2998) and 7/7 NSHP cancers further displayed increased proportions of C>A/G>T transversions (MSI-H vs. non-hypermutated: cell lines 1.5-fold, cancers 2.4-fold) and A>C/T>G transversions (cell line 3.8-fold, cancers 2.0-fold), as well as decreased C>G/G>C transversions (cell lines -13.6-fold, cancers -22.3-fold) (Fig. 2C). Combining cell lines and cancers, this mutator phenotype was significantly associated with the presence of \( POLE \) mutations (NSHP: 90.0%, 9/10 vs. MSI-H: 29.8%, 14/47 vs. non-hypermutated: 1.3%, 3/229; P<0.001). In addition, all 9 NSHP cases with \( POLE \) mutation had at least one missense mutation in the \( POLE \) exonuclease domain (EDM), as compared to only 1/17 non-NSHP cases with \( POLE \) mutation (P<0.001; Supplementary Table S4). 82.4% (14/17) of \( POLE \) mutated non-NSHP samples were MSI-H, and there was no evidence that the non-EDM \( POLE \) mutations in these cases modified the MSI-H phenotype with similar SNV frequencies and transition/transversion spectra compared to MSI-H cancers without \( POLE \) mutation (P>0.05 for all comparisons).

A single NSHP cell line, HT55, exhibited a distinct bias to A>T/T>A transversions (49.0% vs. non-hypermutated mean 4.6%) and was wild-type for \( POLE \), but a similar
case was not present among the TCGA cancers (Fig. 2C). This cell line harbored a heterozygous truncating mutation in the MMR gene \textit{PMS1} which may be related to this mutator phenotype, although Pms1-deficient mice have been reported to show no evidence of tumor predisposition or hypermutation (26).

\textbf{Chromosomal and sub-chromosomal aberrations}

DNA copy-number alterations in cell lines were profiled using Illumina Human 610-Quad BeadChips. As anticipated, DNA copy-number profiles were similar between cell lines and primary cancers (Fig. 3A). Hypermutated MSI-H and NSHP cases showed stable profiles with a modal chromosome copy-number of 2n. In contrast, non-hypermutated cases tended to exhibit unstable profiles with modal chromosome copy-numbers ranging from 2n to 4n in cell lines and 2n to 6n in cancers. For non-hypermutated groups, the most common deleted chromosome arms were 8p, 17p (including \textit{TP53}) and 18q (including \textit{SMAD4}), and the most common gained regions were chromosome 7, 8q (including \textit{MYC}), 13 and 20q. Some differences were apparent between cell lines and primary cancers, including more frequent chromosome 4 deletions in non-hypermutated and chromosome 7 gains and 18q deletions in hypermutated cell lines.

Focal regions significantly targeted by DNA copy-number alterations in both cell lines and cancers were deduced using GISTIC2.0 software (23). Eleven and four significant focal regions were found to overlap between non-hypermutated and
hypermutable groups, respectively (Fig. 3B-C, Supplementary Table S5). For non-hypermutable cases, overlapping regions included gain/amplification of MYC, a known key mediator of colorectal tumorigenesis (27), and deletion of the candidate cancer genes PARK2 and WRN, detected in 29.5% and 36.3% of cell lines and 21.9% and 31.6% of primary cancers, respectively. The four recurrent regions identified in hypermutated cases – which were also present in the non-hypermutable group – were all known fragile sites (including FHIT, A2BP1, WWOX or MACROD2), the functional relevance of which is uncertain.

**Cancer genes and pathways**

Gene mutation profiles, excluding silent mutations, were compared between CRC cell lines and primary cancers. Only genes with well-defined expression in cell lines (reads per kilobase per million reads mapped (RPKM) >1 in at least 3 out of 13 CRC cell lines analyzed by RNA-Seq) were considered (n=20,702 genes; Supplementary Table S6). NSHP cases were excluded from this comparison due to the small sample size.

Mutation landscapes of cell lines markedly resembled those of primary cancers. In both cohorts, non-hypermutable cases displayed a small number of distinct mutation peaks, with APC, TP53 and KRAS being the most common targets, while hypermutated MSI-H cases showed frequent mutations in multiple genes, with the anticipated bias to genes comprising nucleotide repeats (Supplementary Tables S7). Significant overlap
was observed for the top 5% of mutated genes (based on the proportion of affected samples), with 54 and 62 genes intersecting for non-hypermutated and MSI-H groups, respectively (P<0.001, Supplementary Tables S8). Overlapping genes included the expected members of the WNT, MAPK, PI3K, TGF-β and p53 pathways. For non-hypermutated cases, these comprised APC, CTNNB1, FBXW7, KRAS, BRAF, PIK3CA, SMAD4 and TP53, as well as the candidate cancer genes, AXIN2, BCL9L, FAT1, SOX9, ERBB3, PIK3C2B, TIAM1 and ATM. For MSI-H cases these encompassed APC, FBXW7, PIK3CA and TGFBR2, along with the candidates CREBBP, TCF7L2, RNF43 and ACVR2A. Mutation distributions across CRC-associated signaling pathways were also similar between non-hypermutated cell lines and cancers, with the same pathway members tending to show the highest mutation frequencies (Fig. 4). Greater variability in mutation frequencies between pathway members was observed for MSI-H cases, at least partly related to the higher mutation background and smaller sample size. A notable difference for MSI-H cases was a differential prevalence of CTNNB1 mutations, which were frequent in cell lines (47%) but not reported in primary cancers (P<0.001).

In addition to the established CRC-associated pathways, significant enrichment was observed for mutations in chromatin-state regulators (P<0.001; Gene Ontology Enrichment Analysis). These comprised ASH1L, CHD6, PRDM2 and MLL3 in non-hypermutated and ARID1A, EP300, EP400, MLL2, CHD6, PRDM2, TRRAP and SRCAP in MSI-H cases (Supplementary Table S8). Approximately 49% and 19% of non-hypermutated and 100% and 93% of MSI-H cell lines and primary cancers carried mutations across these candidates, respectively.
Integrating mutation and DNA copy-number data across cell lines and primary cancers showed the anticipated tumor suppressor signatures for APC, TP53, SMAD4 and SOX9 with a significant over-representation of two mutational hits (2+ mutations or 1 mutation and loss of heterozygosity; P<0.023). There was further an association of mutation in BRAF, ERBB3, PIK3CA and KRAS with copy-number gain of ≥5 at the respective chromosomal regions (P<0.018). Trends to mutual exclusivity between pathway member mutations were observed for KRAS and BRAF, TP53 and ATM (P<0.005).

**Mutation and DNA copy-number differences in paired CRC cell lines**

Included in our CRC cell line panel were five pairs/triplets originally derived from the same tumor (COLO201/COLO205, CX-1/HT29, Gp2D/Gp5D, LS174T/LS180, DLD1/HCT8/HCT15), and one pair derived from a primary tumor and subsequent lymph node metastasis (SW480/SW620). LS174T and LS180 were established by trypsin treatment and scraping of primary cultures from the same tumor, respectively (28), and have been shown to differ with respect to E-cadherin expression and cell adhesion, with LS174T displaying complete loss of E-cadherin protein (2, 29).

Assessment of the overlap between mutations detected in paired cell lines identified substantial discrepancies, with 63 and 356 mutational differences in the non-hypermutated pairs COLO201/COLO205 and CX-1/HT29, and 2763, 480, 6503, 5377 and 6369 mutational differences in the MSI-H pairs Gp2D/Gp5D, LS174T/LS180,
DLD1/HCT8, DLD1/HCT15 and HCT8/HCT15 (Fig. 5). 849 mutations differed in the non-hypermutated primary/metastasis pair SW480/SW620. Non-silent and silent mutations contributed in similar proportions to these discrepancies (54.4% vs. 56.6%), suggesting no selection for these differential alterations.

DNA copy-number profiles showed multiple differences for non-hypermutated pairs, with 41.1% and 53.8% of the genome varying for COLO201/COLO205 and CX-1/HT29 (Fig. 5). In contrast, discrepancies were limited in MSI-H pairs with 0.2%, 0.4%, 6.6%, 2.9% and 4.3% of the genome differing between Gp2D/Gp5D, LS174T/LS180, DLD1/HCT8, DLD1/HCT15 and HCT8/HCT15, respectively. 47.5% of the genome differed in the non-hypermutated primary/metastasis pair SW480/SW620.

Notably, mutation differences between paired cell lines did not obscure known driver genes. For the established CRC genes APC, TP53, SMAD4, PIK3CA, KRAS and BRAF, 29 of 33 (87.9%) non-silent mutations were concordant between cell line pairs.

**DISCUSSION**

In this study we show that the mutation and DNA copy-number landscapes determined for 70 CRC cell lines closely resemble those of primary tumors, underscoring the utility of cell lines as an appropriate model system for this malignancy. The three molecular subtypes of CRC recently defined by the TCGA, non-
hypermutated, hypermutated with microsatellite instability and hypermutated without microsatellite instability were faithfully captured in the cell line panel. As expected, MSI-H cell lines exhibited hypermethylation of the MLH1 promoter and/or mutations in MMR genes, while hypermutated lines without microsatellite instability were instead characterized by mutations in the exonuclease domain of POLE. Consistent with defective DNA polymerase ε proofreading function, the latter tumors showed a nucleotide substitution hypermutator phenotype with a bias to C>A/G>T and A>C/T>G transversions. Germline missense mutations in the POLE exonuclease domain have recently been associated with familial predisposition to CRC (30), and elevated tumor predisposition and base-substitution mutations observed in Pole exonuclease-mutant mice (31). Similarly, somatic missense mutations in the POLE exonuclease domain have been reported in 7% of endometrial cancers, with good evidence of associated hypermutation (32). While differences in prognosis and chemotherapy response have been extensively documented for non-hypermutated versus MSI-H tumors (2-4), clinical characteristics of POLE-mutant NSHP tumors are unknown. Our identification of two cell lines representative of this CRC subtype will facilitate investigation of the specific aspects of the biology of these tumors, particularly in relation to identifying therapeutics which may exploit their unique genomic instability.

Consistent with previous lower-resolution data (33), DNA copy-number profiles were similar between CRC cell lines and primary tumors, with non-hypermutated cases tending to exhibit chromosomal instability, while both MSI-H and NSHP cases had overall stable copy-number profiles. Patterns of whole and partial chromosome gains
and losses were largely concordant. Besides recurrent gain/amplification of the established CRC gene \textit{MYC}, recurrent regions of deletion in non-hypermutated cases included the candidate cancer genes \textit{PARK2} and \textit{WRN}. \textit{WRN} has important roles in homologous recombination repair, \textit{MUTYH}-mediated repair of oxidative DNA damage and telomeric recombination (34-36), and \textit{WRN} germline mutations are associated with chromosomal instability and cancer predisposition (37). \textit{PARK2} deletion has been previously reported in sporadic CRC, and \textit{Park2} deficiency shown to accelerate intestinal adenoma development in \textit{Apc} mutant mice (38).

The mutation landscapes in non-hypermutated and hypermutated MSI-H cases showed close resemblance in cell lines and primary tumors, with the expected alterations in the WNT, MAPK, PI3K, p53 and TGF\(\beta\) pathways. Besides these main CRC-associated pathways, multiple chromatin-state regulators were enriched among the top 5\% of mutated genes, including proteins involved in chromatin remodeling (\textit{ARID1A}, \textit{CHD6}, \textit{SRCAP}) and histone methylation or acetylation (\textit{ASH1L}, \textit{EP300}, \textit{EP400}, \textit{MLL2}, \textit{MLL3}, \textit{PRDM2}, \textit{TRRAP}). Overall, \~24\% of non-hypermutated and \~96\% of MSI-H cases harbored mutations in these genes. Trends for mutations to cluster in chromatin regulators have been reported in multiple other cancer types including liver cancers (39), gastric adenocarcinoma (40) and transitional cell carcinoma of the bladder (41), suggesting potential pathogenicity of these alterations.

Paired cell lines originating from the same tumor showed considerable differences for both mutations and DNA copy-number profiles. Non-hypermutated pairs
differed for up to 356 mutations and ~54% of genome copy-number, while MSI-H pairs differed for up to 6503 mutations and ~6.6% of genome copy-number. The two main possible reasons for these differences are pre-existing heterogeneity between the cells from the original tumor grown out to establish these paired lines, or acquisition of alterations as a result of long-term culture. Assuming a normal mutation rate of $10^{-8}$ (per bp per cell generation) for non-hypermutated and 100-fold elevated rate for hypermutated pairs and absence of selection (42), cell lines established from tumor cells separated by an average of 328 and 66 replications would be expected to exhibit the observed mutational differences with a >99% probability, respectively (Supplementary Data, Supplementary Table S9). In contrast, mutations acquired during serial passaging in culture are not anticipated to reach a detectable 10% level prior to 9,034 and 4,266 replications, respectively (Supplementary Data, Fig. 6). Although serial cell passaging may have been common-place at the time when these paired cell lines were established >20 years ago, stock-keeping practices to limit the number of replications were soon introduced. Pre-existing mutation heterogeneity in the original tumor therefore seems highly likely to account for the majority of the detected differences. Numbers of replications may be expected to be larger between cells from a primary tumor and subsequent metastasis, and accordingly our corresponding non-hypermutated cell line pair showed a ~4-fold higher number of mutational differences than the other non-hypermutated pairs. Importantly, mutational differences across paired cell lines did not obscure known driver genes, with ~88% of non-silent mutations in established CRC genes concordant between pairs.
Despite the high level of similarity between CRC cell lines and primary tumors, there were a number of differences. These included overall higher mutation and DNA copy-number frequencies, as well as a greater proportion of detected InDels in cell lines. These discrepancies may be related to differences in exome-capture and sequencing platforms, bioinformatics pipelines, the presence of contaminating normal tissue in primary cancers and accuracy of assigning somatic alterations in cell lines. Cell lines further showed a higher proportion of hypermutated cases, and exhibited differences in the prevalence of aberrations for certain genes and genomic regions such as CTNNB1 mutations in MSI-H cases. These latter findings may be a reflection of preferential growing out of cell lines from primary tumors (or their respective subclones) that contain these aberrations, a contention supported by the observation that only ~10-15% of CRCs give rise to cell lines (43). Another possibility is that some of the mutations have been acquired and selected for in tissue culture.

A caveat to our analysis of protein-coding genes is that we could not report on untranslated exonic regions (UTRs), as the latter were inconsistently covered by our study and the TCGA. UTR mutations can impact on RNA splicing, stability or translation as previously highlighted for MSI-H CRC (44). In the comparison of gene mutation profiles we chose to exclude silent mutations (other than those affecting splice-sites), although some of these may similarly have functional consequences (45).

In conclusion, our comparative analysis of the genomic landscapes of human CRC cell lines and TCGA-analyzed primary cancers identified cell lines representative
of the three main mutational CRC subtypes. Within these molecular subtypes, while some differences were evident, cell lines showed globally similar genetic alterations to primary cancers including genome-wide mutation, DNA copy-number and driver gene mutation profiles. Accordingly, gene expression profiles of CRC cell lines have previously been shown to broadly represent those of primary tumors (46). Our genomic data significantly expand on cancer cell line characterization efforts by the major cancer genome centers, such as the Cancer Cell Line Encyclopedia project which currently reports mutation data for 1,500 selected genes on 62 CRC cell lines (5). Our data will help to inform investigations of the molecular basis of CRC pathogenesis, inherent and acquired drug resistance, and exploration of novel treatment modalities for this malignancy.

ACKNOWLEDGMENTS

We thank the Victorian Cancer BioBank for patient specimens and Prof. Manfred Schwab at the DKFZ for providing cell lines.
REFERENCES


32. Church DN, Briggs SE, Palles C, Domingo E, Kearsey SJ, Grimes JM, et al. DNA polymerase \{varepsilon\} and delta exonuclease domain mutations in endometrial


FIGURE LEGENDS

Fig. 1. Mutation frequencies in 70 human CRC cell lines. Cell lines separate into distinct hypermutated and non-hypermutated cases. MSI-H, microsatellite instability-high; CIMP, CpG island methylator phenotype; MLH1 meth., MLH1 promoter hypermethylation, MMR mut., DNA mismatch repair gene mutation; Inset, mutations in MMR genes among the hypermutated samples, with cases in the same order as in the main graph.

Fig. 2. Types of mutations for 70 human CRC cell lines and 223 TCGA-analyzed primary cancers. (A, B) Counts of SNVs and InDels, and (C) proportions of nucleotide transitions and transversions. Paired cell lines originating from the same tumor are indicated by identical colored stars. MSI-H, microsatellite instability-high; NSHP, nucleotide-substitution hypermutator phenotype; POLE EDM, missense mutation in the POLE exonuclease domain.

Fig. 3. Genome-wide DNA copy-number aberrations in 63 unique CRC cell lines and 213 TCGA primary cancers stratified into non-hypermutated and hypermutated cases. (A) Absolute DNA copy-number gains and losses relative to diploid status (2n) deduced using oncoSNP. (B, C) Minimal significant regions of chromosomal loss (blue) and gain/amplification (red) deduced using GISTIC2.0. Selected candidate genes in overlapping regions are indicated and total numbers of genes are shown in brackets.
Fig. 4. Mutation frequencies for WNT, MAPK, PI3K, TGF-β and p53 pathway members as well as chromatin regulators identified among the top 5% of mutated genes overlapping between CRC cell lines and TCGA-analyzed primary cancers for non-hypermutated or hypermutated MSI-H samples.

Fig. 5. Overlap of mutations and DNA copy-number states between paired CRC cell lines. COLO201/COLO205, CX-1/HT29, Gp2D/Gp5D, LS174T/LS180 and DLD1/HCT8/HCT15 were derived from the same tumor, SW480/SW620 from a primary tumor and subsequent lymph node metastasis.

Fig. 6. Simulation of the acquisition of mutations in cell culture. The process of serial-passage was modelled with cell numbers repeatedly increasing from $1 \times 10^5$ to $2 \times 10^6$ cells. Proportions of cells containing the most frequent mutation by passage number for five simulations, using fixed mutation probabilities of (A) $10^{-8}$ per base per cell replication for non-hypermutated and (B) $10^{-6}$ for hypermutated MSI-H tumors. The black diagonal trendline is the least squares fit on the log-log scale and the small vertical lines are the 99% prediction intervals. The horizontal red line corresponds to a sequencing mutation detection threshold of 10%.
Figure 1
Figure 5
Figure 6
Colorectal cancer cell lines are representative models of the main molecular subtypes of primary cancer

Dmitri Mouradov, Clare Sloggett, Robert N Jorissen, et al.

Cancer Res Published OnlineFirst April 22, 2014.

Updated version
Access the most recent version of this article at:
doi:10.1158/0008-5472.CAN-14-0013

Supplementary Material
Access the most recent supplemental material at:
http://cancerres.aacrjournals.org/content/suppl/2014/04/22/0008-5472.CAN-14-0013.DC1

Author Manuscript
Author manuscripts have been peer reviewed and accepted for publication but have not yet been edited.

E-mail alerts
Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions
To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions
To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.