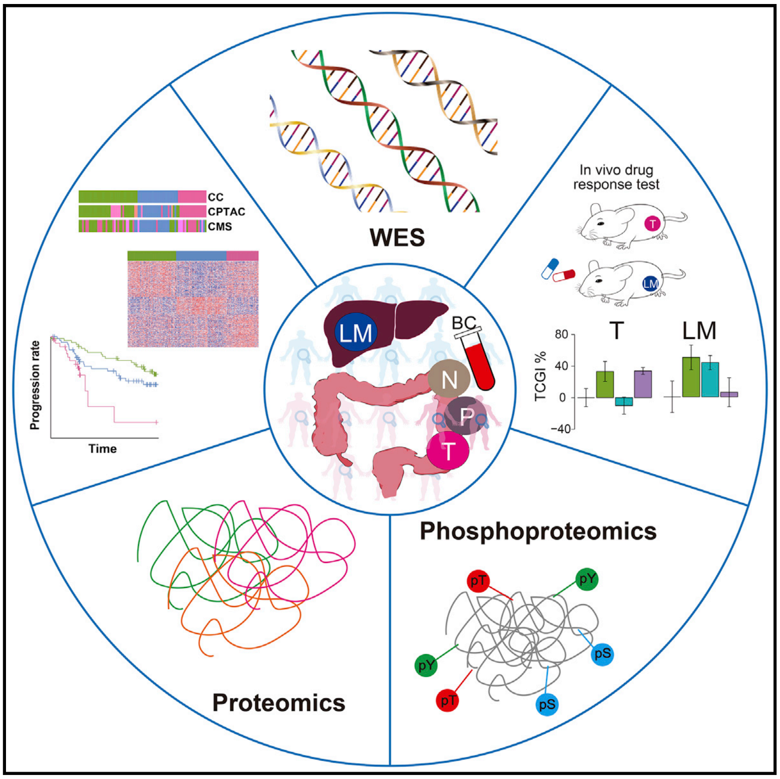
Cancer Cell

Integrated Omics of Metastatic Colorectal Cancer

Graphical Abstract



Highlights

- A large-scale proteogenomics study of metastatic colorectal cancers
- Phosphoproteomic pattern distinguishes metastasis and predicts drug response
- A workflow from generation of large omics datasets to *in vivo* drug testing models
- Improves the selection of treatment strategies for patients without druggable mutation

Authors

Chen Li, Yi-Di Sun, Guan-Yu Yu, ..., Jia-Rui Wu, Wei Zhang, Rong Zeng

Correspondence

wujr@sibs.ac.cn (J.-R.W.),
weizhang2000cn@163.com (W.Z.),
zr@sibcb.ac.cn (R.Z.)

In Brief

Li et al. provide a global proteogenomic landscape for metastatic colorectal cancer in a Chinese cohort. Proteomic and phosphoproteomic profiling of primary tumors successfully distinguishes cases with metastasis and, together with network analysis, accurately reflects the drug responses of primary and metastatic tumors.

Li et al., 2020, Cancer Cell *38*, 1–14
November 9, 2020 © 2020 Elsevier Inc.
https://doi.org/10.1016/j.ccell.2020.08.002

CellPress

Please cite this article in press as: Li et al., Integrated Omics of Metastatic Colorectal Cancer, Cancer Cell (2020), https://doi.org/10.1016/j.ccell.2020.08.002

Cancer Cell

CellPress

Article

Integrated Omics of Metastatic Colorectal Cancer

Chen Li,[1,7,10] Yi-Di Sun,[1,8,10] Guan-Yu Yu,[2,10] Jing-Ru Cui,[1,10] Zheng Lou,[2,10] Hang Zhang,[2] Ya Huang,[1,4] Chen-Guang Bai,[9] Lu-Lu Deng,[9] Peng Liu,[2] Kuo Zheng,[2] Yan-Hua Wang,[1,4] Qin-Qin Wang,[1,4] Qing-Run Li,[1] Qing-Qing Wu,[1] Qi Liu,[5] Yu Shyr,[5] Yi-Xue Li,[3,4,6] Luo-Nan Chen,[1,3,4] Jia-Rui Wu,[1,3,4,*] Wei Zhang,[2,*] and Rong Zeng[1,3,4,11,*]

[1]CAS Key Laboratory of Systems Biology, Shanghai Institute of Biochemistry and Cell Biology, Center for Excellence in Molecular Cell Science, Chinese Academy of Sciences, Shanghai 200031, China

[2]Colorectal Surgery Department, Changhai Hospital, Naval Medical University, Shanghai 200433, China

[3]CAS Key Laboratory of Systems Biology, Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Hangzhou 310024, China

[4]School of Life Science and Technology, ShanghaiTech University, Shanghai 201210, China

[5]Department of Biostatistics and Center for Quantitative Sciences, Vanderbilt University School of Medicine, Nashville, TN 37232, USA

[6]Bio-Med Big Data Center, Key Laboratory of Computational Biology, CAS-MPG Partner Institute for Computational Biology, Shanghai Institute of Nutrition and Health, University of Chinese Academy of Sciences, Chinese Academy of Sciences, Shanghai 200031, China

[7]Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China

[8]Institute of Neuroscience, CAS Center for Excellence in Brain Science and Intelligence Technology, Chinese Academy of Sciences, Shanghai 200031, China

[9]Department of Pathology, Changhai Hospital, Naval Medical University, Shanghai 200433, China

[10]These authors contributed equally

[11]Lead Contact

*Correspondence: wujr@sibs.ac.cn (J.-R.W.), weizhang2000cn@163.com (W.Z.), zr@sibcb.ac.cn (R.Z.)

https://doi.org/10.1016/j.ccell.2020.08.002

SUMMARY

We integrate the genomics, proteomics, and phosphoproteomics of 480 clinical tissues from 146 patients in a Chinese colorectal cancer (CRC) cohort, among which 70 had metastatic CRC (mCRC). Proteomic profiling differentiates three CRC subtypes characterized by distinct clinical prognosis and molecular signatures. Proteomic and phosphoproteomic profiling of primary tumors alone successfully distinguishes cases with metastasis. Metastatic tissues exhibit high similarities with primary tumors at the genetic but not the proteomic level, and kinase network analysis reveals significant heterogeneity between primary colorectal tumors and their liver metastases. *In vivo* xenograft-based drug tests using 31 primary and metastatic tumors show personalized responses, which could also be predicted by kinase-substrate network analysis no matter whether tumors carry mutations in the drug-targeted genes. Our study provides a valuable resource for better understanding of mCRC and has potential for clinical application.

INTRODUCTION

Colorectal cancer (CRC) is the fourth most deadly cancer worldwide, with almost 900,000 deaths annually (Dekker et al., 2019). Aging, unfavorable diet, and lifestyle all increase the risk of CRC (Kuipers et al., 2015). CRC exhibits high heterogeneity (Allison and Sledge, 2014; Punt et al., 2017), with molecularly defined subgroups that differ in their prognosis. Previous studies in TCGA and CPTAC colorectal cohorts have characterized multiomic features and molecular heterogeneity (Cancer Genome Atlas, 2012; Guinney et al., 2015; Vasaikar et al., 2019; Zhang et al., 2014). However, these studies focus more on non-metastatic states in non-Asian populations (Imperiale et al., 2018; Murphy et al., 2019; Simon et al., 2011; Tawk et al., 2015).

Despite the increasing advances in treatment, mortality from CRC, especially from metastatic CRC, remains high among cancer-related deaths (Bray et al., 2018; Dekker et al., 2019). Currently, only DNA mismatch-repair status, RAS mutation, and BRAF mutation status influence clinical decision-making (Punt et al., 2017). Multi-omic characteristics anticipated to contribute to improving therapy will thus lead to precise and individualized care (Kuipers et al., 2015). The combination of genomics and proteomics can provide additional insights, which may not be deciphered by genomic analysis alone (Bhullar et al., 2018; Wu et al., 2019).

Here, we present a global integration of genomics, proteomics, and phosphoproteomics in metastatic CRC from a Chinese cohort. Integrated analyses of multi-omic data demonstrated distinct proteomic and phosphoproteomic characteristics for subtypes and metastasis. Kinase-substrate correlations were identified as accurate indicators of drug response for potential treatment. Collectively, the results of our study provide a rich resource that contains promising targets and therapeutic assessments for CRCs.

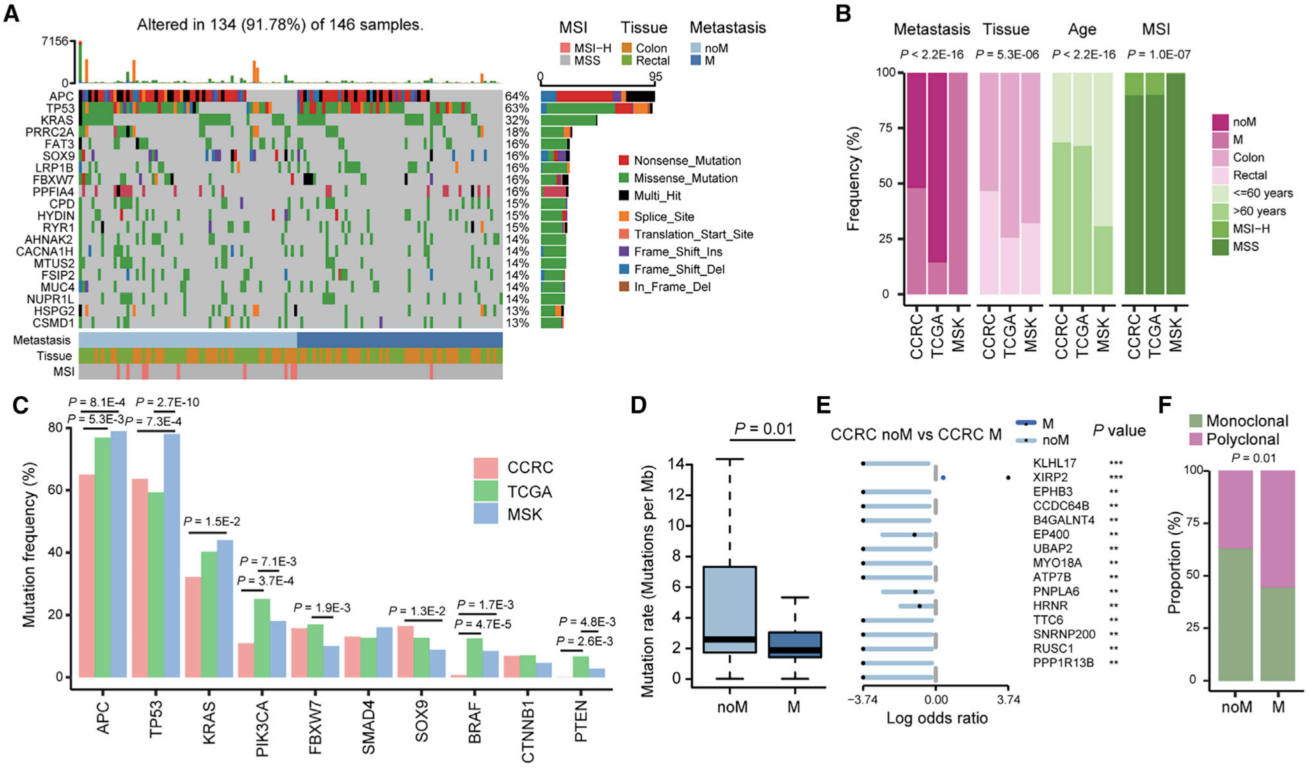


Figure 1. Mutation Landscape of CCRC

(A) Genetic profile and associated clinicopathologic features of all 146 CCRC patients. Bar plot on the top indicates total number of somatic mutations in each patient. Bar plot at the right represents the distribution and compositions of mutation types in each gene.

(B) Comparison of clinical characteristics among CCRC, TCGA, and MSK cohorts (Yaeger et al., 2018). The p values were calculated by chi-square test.

(C) Comparison of frequently mutated genes between CCRC, TCGA, and MSK cohorts. The p values were calculated by Fisher's exact test.

(D) Comparison of mutation burden between non-mCRC and mCRC. The p value was calculated by two-tailed Student's t test. The line and box represent median and upper and lower quartiles, respectively.

(E) Differentially mutated genes between non-mCRC and mCRC. The p values were calculated by Fisher's exact test. **p < 0.01; ***p < 0.001.

(F) Clonal differences between non-mCRC and mCRC. The p value was calculated by chi-square test. M, mCRC; noM, non-mCRC. See also Figures S1–S3 and Table S1.

RESULTS

Overview of the Study

We applied multi-omics-based profiling to 146 Chinese patients with CRC (CCRC) from Changhai Hospital (Shanghai, China), including 70 metastatic CRC (mCRC) patients and 76 earlier-stage non-metastatic CRC (non-mCRC) patients. Primary tumor tissues (T), remote normal tissues (N), para-carcinoma tissues (P, normal adjacent tissues), and matched peripheral blood cells (BC) were obtained from each patient. For mCRC, 43 available distant liver metastatic tissues (LM) were also studied. We performed whole-exome sequencing (WES) on 330 samples (Figure S1), including paired primary tumor and peripheral BC samples (128 T-BC pairs) or N (18 T-N pairs), and 38 DNA quality-controlled LMs. Hybrid spectral libraries of CCRC proteome or CCRC phosphoproteome were generated by MaxQuant and Spectronaut as described in the STAR Methods. The hybrid CCRC proteome spectral library included 179,382 precursors, 113,291 peptides, 11,510 protein groups, and 9,942 gene products. The hybrid CCRC phosphoproteome spectral library included 116,121 phosphoprecursors, 65,851 phosphopeptides, 9,977 phosphoprotein groups, and 7,125 phosphogene products. The proteomes (8,450 quantified protein groups) and phosphoproteomes (47,786 quantified phosphosites) of 480 samples consisting of 145 paired T-N-P tissues, a pair of T-N tissues, and 43 LM tissues, were characterized using data-independent acquisition methods (Figures S1 and S2). These patients had a median follow-up time of 1,240 days (Table S1).

Mutational Landscape of Chinese CRCs

A median of 107 non-synonymous, somatic, single-nucleotide variants and 7 insertions or deletions were identified in primary tumors of 146 CRC patients (Table S1), similar to the results obtained for the TCGA CRC cohort. The most frequently observed cancer-associated mutations in this cohort were *APC* (65% mutation frequency), *TP53* (64%), and *KRAS* (32%) (Figure 1A), consistent with previous studies. The clinicopathological characteristics distinguished the CCRC cohort from the TCGA (Cancer Genome Atlas, 2012) or MSK CRC cohort (Yaeger et al., 2018) (Figure 1B). Compared with the TCGA CRC dataset, CCRC had a higher ratio of mCRC patients (47.9% versus 14.4%; p < 2.2E−16; Figure 1B). In addition, CCRC contained the highest proportion of rectal cancer cases of the three cohorts (46.6% versus 25.5% versus 31.7%; p = 5.3E−06; Figure 1B), consistent with previous studies that showed a high incidence of rectal cancers among Asian populations (Deng, 2017; Sung et al., 2019).

We found that the *APC* mutation frequency was significantly lower in CCRC compared with the other two cohorts (65.1% versus 76.9%; p = 5.3E−03 for TCGA and 65.1% versus 79%; p = 8.1E−04 for MSK), as was the mutation frequency of *TP53* compared with that of the MSK cohort (63.7% versus 78%; p = 7.3E−04). The mutational hotspots in *APC* and *TP53* in CCRC were similar to those in the TCGA cohort (Figure S3A). Furthermore, frequencies of *BRAF* and *PTEN* mutations in CCRC were also significantly lower than in the Western datasets (Figure 1C and Table S1). In contrast, we found higher mutation frequencies in the *PRRC2A*, *PPFIA4*, *CPD*, and *NURP1L* genes in CCRC (Figure S3B). Considering that the CCRC cohort was demographically distinct from the TCGA cohort, and consisted of more advanced cases and more cases of rectal disease (Figure 1B), we performed propensity score matching (PSM) of clinical characteristics between the two cohorts for genomic feature comparison (STAR Methods; Figure S3C). These mutations were preferentially found in the CCRC cohort after the PSM (Figures S3D and S3E), indicating genetic signatures potentially unique to Asian CRCs.

To identify genes associated with metastasis, we compared the frequencies of genomic alterations in non-mCRC and mCRC from the CCRC cohort. Mutation burdens in primary tumors of mCRC were decreased compared with those of non-mCRC (Figure 1D), which is also observed in non-hypermutated CRC cases (Figure S3F). Among the most frequently mutated genes in CRC (Figure 1C), only *SMAD4* showed a significantly higher mutation rate in primary tumors of mCRC patients (20% versus 6.6%; p = 0.015; Table S1); *XIRP2* was also significantly highly enriched in primary tumors of mCRC patients (Figure 1E), which has been reported to correlate with breast cancer progression (Kroigard et al., 2018).

We further applied non-negative matrix factorization to extract mutational signatures. Four signatures were revealed in the 146 CCRC primary tumors, where COSMIC SBS 6, SBS 1, SBS 45, and SBS 5 were identified as defined previously (Figures S3G and S3H). Notably, the contribution rate of SBS 1 to mCRC was significantly higher than that to non-mCRC (Figure S3I), thereby demonstrating a more severe endogenous mutation status in mCRC. Moreover, genes enriched for SBS 1, such as *HYDIN*, *C1QB*, and *COL22A1*, have been previously reported as metastatic signatures of colon and breast cancers (Naba et al., 2014; Zhang et al., 2015b) (Figure S3J). The most frequent somatic copy number alterations (SCNAs) showed no evident differences between mCRC and non-mCRC primary tumors (Figures S3K and S3L). However, primary tumors of mCRC patients exhibited a more polyclonal architecture in comparison with that of non-mCRC patients (63% and 39%; p = 0.01; Figure 1F), suggesting the metastatic probability of mCRC in T.

Subtype Classification Based on Proteomic Data

We performed consensus clustering of 2,440 differentially expressed proteins between primary tumors and N (STAR Methods), in order to explore whether our deep proteomes can provide insight into cellular and molecular heterogeneity associated with CRC (Table S2). Among the 146 CCRC primary tumors, three consensus clusters (CCs) were identified (Figures 2A and S4A). CC1 was characterized by increased RNA processing and DNA mismatch repair (MMR). Upregulated proteins in CC2 were enriched for extracellular matrix (ECM)-receptor integration, focal adhesion, and immune-related pathways. CC3 featured enrichment for both the upregulation of DNA replication and metabolic pathways.

Clinicopathologic characteristics showed no significant differences among different subtypes except for pre-surgery treatment (p = 0.014; Fisher's exact test), probably due to the slight enrichment of mCRC in CC2 and CC3. The three subtypes had different relapse-free survival probabilities (p = 0.014; Figure 2B), while subtyping remained an independent prognostic factor after adjusting for tumor stage and pre-surgery treatment by multivariate analysis (p = 0.017; Figure S4B). In addition, mCRC patients in CC3 also showed the worst probability of relapse-free survival compared with mCRC patients in CC1 and CC2 subtypes (p = 0.004; Figure 2C). By contrast, non-mCRC patients showed no significant difference in relapse-free survival among the three subtypes (Figure S4C), possibly due to the overall good prognosis for non-metastatic CRC through various treatments. In addition, all mCRC primary tumors across the different CC subtypes showed distinct characteristics (Figure S4D). Specifically, the proteins elevated in CC3 were primarily associated with the citrate cycle, oxidative phosphorylation, and metabolic pathways (Figure S4E).

We then identified differentially mutated genes or SCNAs among the three subtypes (see STAR Methods; Figure 2D and Table S2). For genes showing differences in mutation, SCNA, and protein levels among the three CC subtypes, we found that mutations were significantly enriched in CC3 and rectal cancers (Figure 2E), consistent with the tumor-location bias of CRC in Asian populations. Next, we found that *FBXW7*, *HYDIN*, and *HSPG2* were significantly enriched in subtype CC3 (Figure S4F). The SCNAs were significantly deleted and proteins were upregulated in the CC3 subtype of the overlapping genes (Figures 2D, 2E, and S4F). We then examined the correlations of the 295 genes with differential SCNAs and protein abundance (Figure 2D). Interestingly, most SCNA genes showed greater deletions in CC3 compared with CC1 and CC2, while proteins in CC3 exhibited higher expression levels (Figures S4G and S4H). These genes were significantly enriched in oxidative phosphorylation, RNA splicing, neutrophil degranulation, and alternative-splicing-related pathways (Figure S4I). Specifically, SCNA genes in the 19q region were deleted with gradually increasing frequency from CC1 to CC3, although the abundances of proteins in this region showed the lowest and highest expression levels in CC2 and CC3, respectively (Figure S4H). For example, *COX6C*, associated with oxidative phosphorylation, and *PTBP1* and *ELAVL1*, in the alternative splicing pathway, all had lower copy numbers in CC3, although their protein levels were significantly higher in CC3 compared with CC1, CC2, and N (Figure S4J). In addition, high protein abundance was correlated with poor probability of relapse-free survival for all three genes (Figure S4J), suggesting that copy number and protein abundance are decoupled in poor-prognosis tumors (Myhre et al., 2013; Zhao and Jensen, 2009).

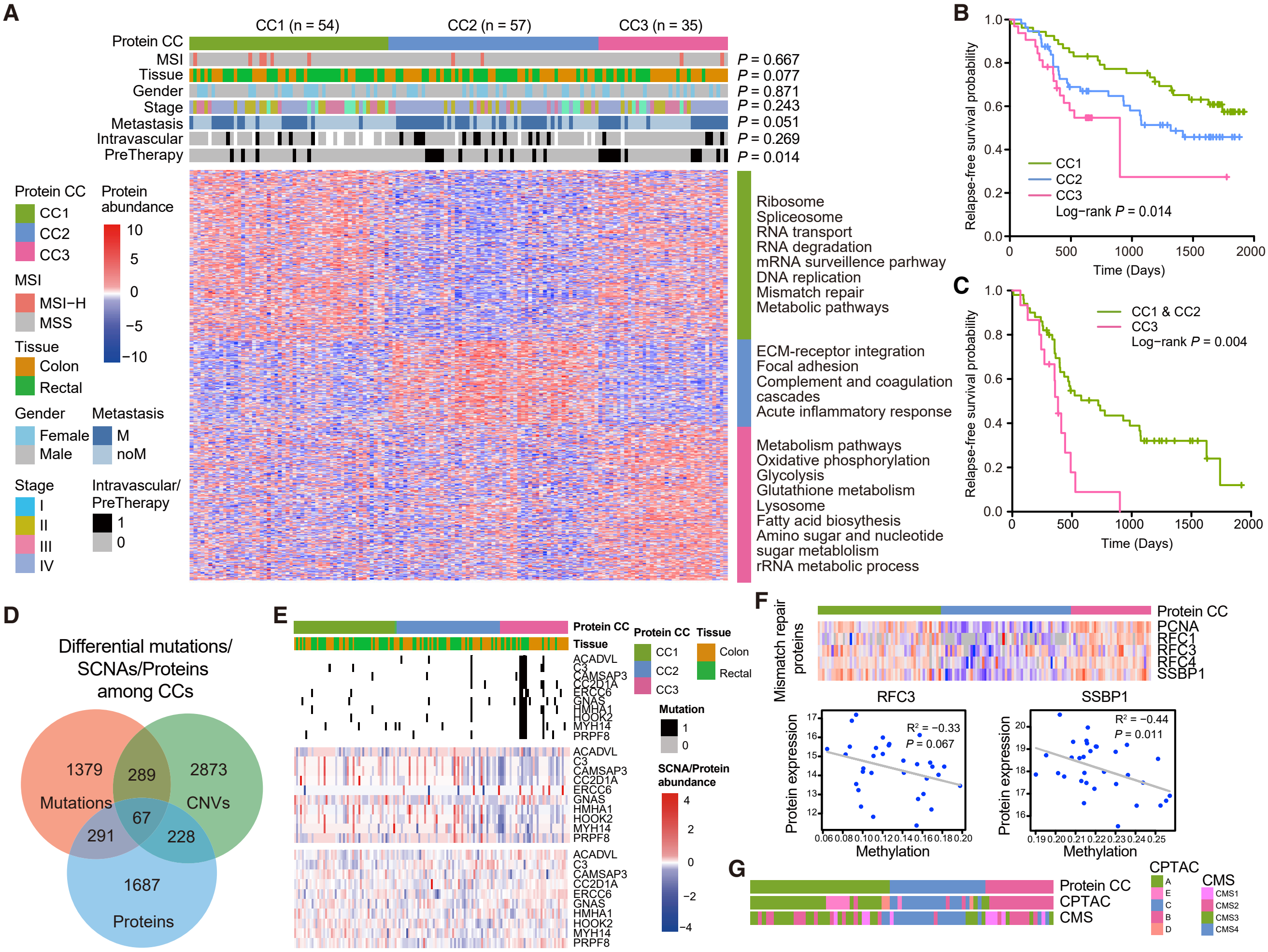


Figure 2. Proteomic Subtyping of CCRC and Clinical Implications for Each Subtype

(A) Consensus clustering based on differentially expressed proteins between tumor and remote normal tissues. Each column represents a patient sample and rows indicate proteins.

(B) Kaplan-Meier curves for relapse-free survival based on proteomic subgroups. The p value was calculated by log rank test.

(C) Kaplan-Meier curves for relapse-free survival based on proteomic subgroups for mCRC. The p value was calculated by log rank test.

(D) Venn diagram illustrates the overlap of differential gene mutations, SCNAs, or proteins among three CCs.

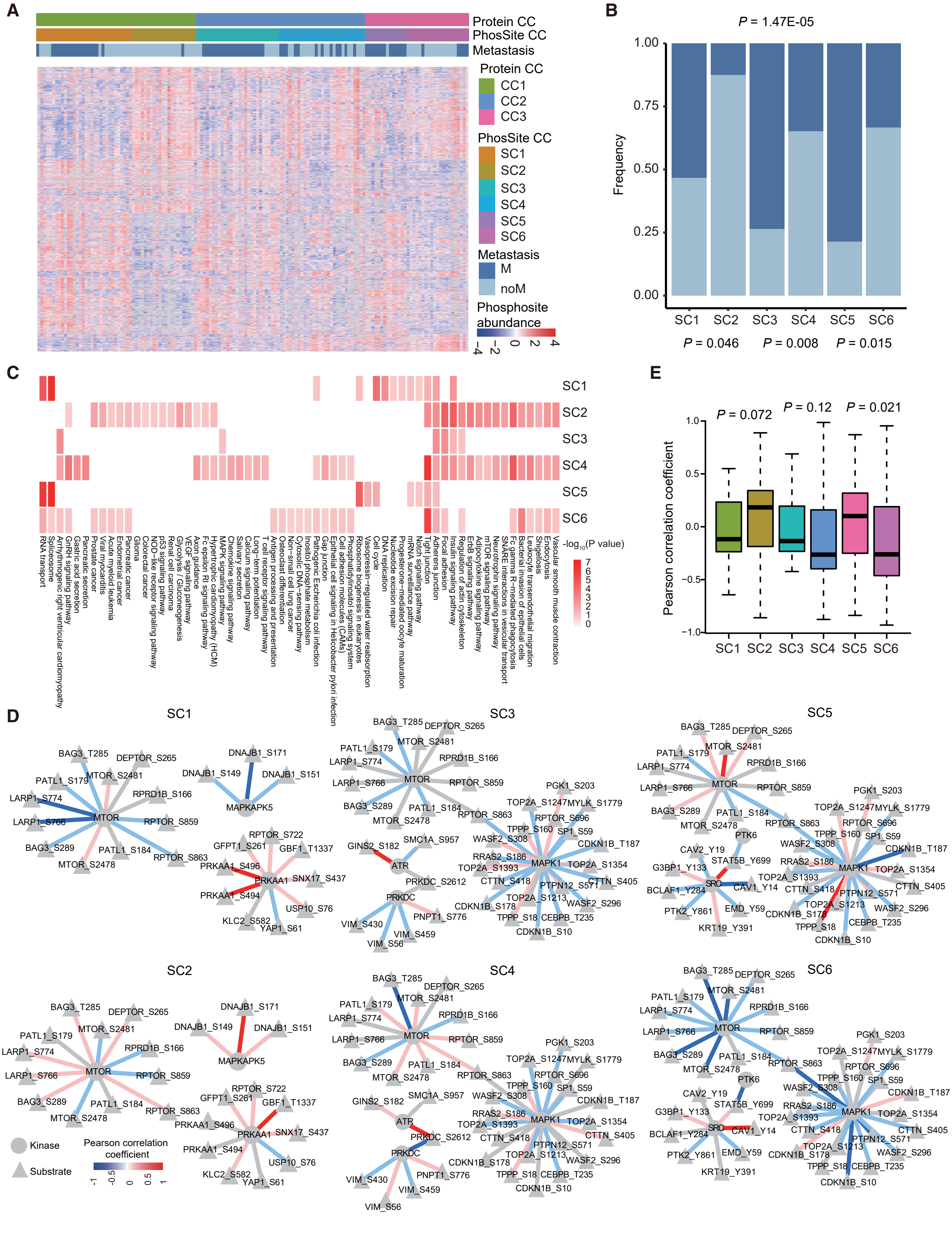
(E) The top 10 differentially mutated genes that also showed differences in SCNA and protein levels.

(F) The expression of proteins enriched in the mismatch repair pathway (top). The correlation between methylation level and protein expression of *RFC3* and *SSBP1* (bottom). Correlation coefficients and p values were calculated by the Spearman correlation method.

(G) Comparison of proteomic subtyping of non-mCRC with previous subtyping results based on RNA (Guinney et al., 2015) or Western CRC patients (Vasaikar et al., 2019; Zhang et al., 2014). M, mCRC; noM, non-mCRC.

See also Figures S1, S2, and S4 and Table S2.

Downregulated proteins in CC2 were enriched for the DNA MMR pathway (Figures 2A and S4K) but showed no apparent correlation with microsatellite instability (MSI) status (Figure 2A). Instead of the commonly reported MMR proteins MLH1, MLH3, MSH2, MSH3, and MSH6, we found that PCNA, RFC1, RFC3, RFC4, and SSBP1 MMR pathway proteins were differentially enriched and downregulated in CC2 compared with other subtypes (Figures 2F and S4L). To further explore the mechanism driving MMR, we selected 32 samples (6, 13, and 13 samples for CC1, CC2, and CC3, respectively), based on protein expression levels, for Illumina 850K methylation array (STAR Methods) and found that the methylation levels of *RFC3* and *SSBP1* were significantly negatively correlated with the expression of their encoded proteins (Figures 2F and S4M). These results revealed that the CC2 subtype bears a unique and non-canonical epigenetic feature correlated with protein pattern. The subtypes identified by proteomic data were generally consistent with the subtyping from previous studies of the CPTAC proteome dataset and CMS classification (Figures 2G and S6B) (Guinney et al., 2015; Zhang et al., 2014). Notably, CC1 matched CPTAC subtypes A and E, containing CMS1 and CMS3, respectively, and showed relatively good prognoses as reported (Guinney et al., 2015; Zhang et al., 2014). CC2, which similarly matched subtype C and CMS4 (Figure 2G), showed typically worse prognosis than



(legend on next page)

Please cite this article in press as: Li et al., Integrated Omics of Metastatic Colorectal Cancer, Cancer Cell (2020), https://doi.org/10.1016/j.ccell.2020.08.002

CC1 (Figures 2A and 2B), due to ECM enrichment with mesenchymal features. The CC3 proteome corresponded to CPTAC subtype B, with enrichment for MSI (Zhang et al., 2014; Vasaikar et al., 2019). However, we did not observe enrichment for MSI in CC3. In our analyses, CC3 matched diverse CMSs (Figure 2G), and demonstrated the worst relapse-free survival probability (Figure 2B), illustrating the complexity of this subtype. These results may be related to the significantly greater number of metastatic patients in our CCRC cohort than in the TCGA/CPTAC cohorts (Figure 1B), while the divergence between CNV and protein expression further leads to the heterogeneity, and consequently inconsistency, in the CC3 subtype.

Phosphoproteomic Profiles Distinguished mCRC from Non-mCRC

In our proteome subtyping, none of the three subtypes were significantly enriched in mCRC or non-mCRC patients (Figure 2A). We found that, in total, 1,487 phosphosites were differentially expressed in primary tumors and N (Table S3). We applied consensus clustering using these differential phosphosites to identify sub-clusters in each CC (STAR Methods; Figure S5A). Interestingly, phosphoproteomic data distinguished the primary tumors from mCRC and non-mCRC in each proteomic subtype ($p = 1.47E-5$, chi-square test), resulting in classification of six phosphoproteomic subtypes (Figures 3A, 3B, and S5B). SC1, SC3, and SC5 were enriched in mCRC, while SC2, SC4, and SC6 were characteristic of non-mCRC.

We performed functional enrichment analysis and found that phosphoproteins with high expression in SC1, SC3, and SC5 were enriched in focal adhesion and adherens junction pathways (Figure 3C). In contrast, phosphosites upregulated in SC2, SC4, and SC6 showed more similarity in function between subtypes and were enriched in *ERBB2* signaling, endometrial cancer, antigen processing and presentation, and Fc gamma R-mediated phagocytosis pathways (Figure 3C). Specifically, phosphosites of MHC1 predicted to participate in the antigen processing and presentation pathway were downregulated in SC1, SC3, and SC5 (Figures S5C and S5D), suggesting an inhibition of T cell activation in the progression of metastasis. Moreover, upregulated phosphosites in the mTOR signaling and glycolysis pathways (Figures S5D and S5E) provided pharmacological insights into CRC metastasis. Based on kinase-substrate relationships from PhosphoSitePlus (Hornbeck et al., 2015), we found mostly negative correlations between kinases and phosphosites in SC1, SC3, and SC5, while positive correlations were increased in SC2 and SC4 (Figure 3D). Specifically, SC6 was characterized by negative regulation between kinases and substrates, suggesting that the non-mCRC in CC3 was more like mCRC with poor prognosis (Figure 3E). Since the phosphosite abundance is potentially affected by either protein expression level or phosphorylation activity, an alternative method is to normalize phosphorylation with protein abundance. This analysis showed that phosphoproteomic data could also distinguish between primary tumors of mCRC and non-mCRC in CC3 (Figures S5F and S5G).

Proteogenomic Characteristics of Metastatic Tumors

For the mCRC, a high concordance between mutational profiles was observed for primary and metastatic tumors (Figures 4A and 4B; Table S4), regardless of the MSI status (Figure S6A). Although primary tumors tended to have more unique mutations, no obvious differences were observed in mutations among the putative driver genes or with mCRC datasets from previous studies (Yaeger et al., 2018) (Figure 4C and Table S4). In addition, the most frequently mutated SCNAs showed no difference between primary and metastatic tumors (Figure 4D), whereas metastatic mCRC tumors exhibited a greater monoclonal proportion compared with primary tumors (60% and 40%; $p = 0.02$; Figure 4E). Together, these results suggested that the metastatic tumors were derived from the primary tumors or from the same ancestral clones.

However, proteomic profiling of metastatic tumors showed obvious differences from that of primary tumors (Figure S6B). Specifically, metastatic tumors had more upregulated proteins compared with normal tissue than did primary tumors (Figure 4F and Table S4). These differentially expressed proteins clearly distinguished metastatic tissues from primary tissues (Figure 4G). Proteins upregulated in metastatic tumors were correlated with ECM-receptor interaction, drug metabolism, focal adhesion, and tight junction (Figure 4H), while proteins downregulated in metastatic tumors were enriched in metabolic pathways, fatty acid degradation, citrate cycle, and oxidative phosphorylation (Figure 4H). In each subtype, proteins that differed in expression from normal tissues were also distinct between primary and metastatic tissues (Figure S6C). For example, proteins upregulated in primary tumors were enriched in leukocyte transendothelial migration, as well as in complement and coagulation in CC2, but citrate cycle and PPAR signaling pathways in CC3 (Figure S6C).

Phosphosite-to-Protein Co-variation in Multiple Tissues of mCRC

Next, we focused on 42 mCRC cases with all four tissue types (N, P, T, and LM). Among them, 10, 21, and 11 cases belonged to CC1, CC2, and CC3, respectively. In each case, we computed the Pearson's correlation coefficients of all four tissues between each two matched pairs of phosphosite abundances versus protein abundances and obtained an array of correlation coefficients for the 42 mCRC cases (STAR Methods). We found that

Figure 3. Phosphoproteomic Profiling in CCRC

(A) Consensus clustering of phosphoproteomic data based on the proteomic subtyping.

(B) The distribution of mCRC and non-mCRC in each phosphoproteomic subtype. The p values were calculated by chi-square test.

(C) Functional enrichment for significant genes in each phosphoproteomic subtype.

(D) Phosphoproteomic regulation networks in the six phosphoproteomic subtypes. The edges represent Pearson's correlation coefficient between kinases and the corresponding substrates.

(E) Distribution of Pearson's correlation coefficients of kinase-substrate networks in (D). The p values were calculated with two-tailed Student's t test. The line and box represent median and upper and lower quartiles, respectively. M, mCRC; noM, non-mCRC.

See also Figures S1, S2, and S5 and Table S3.

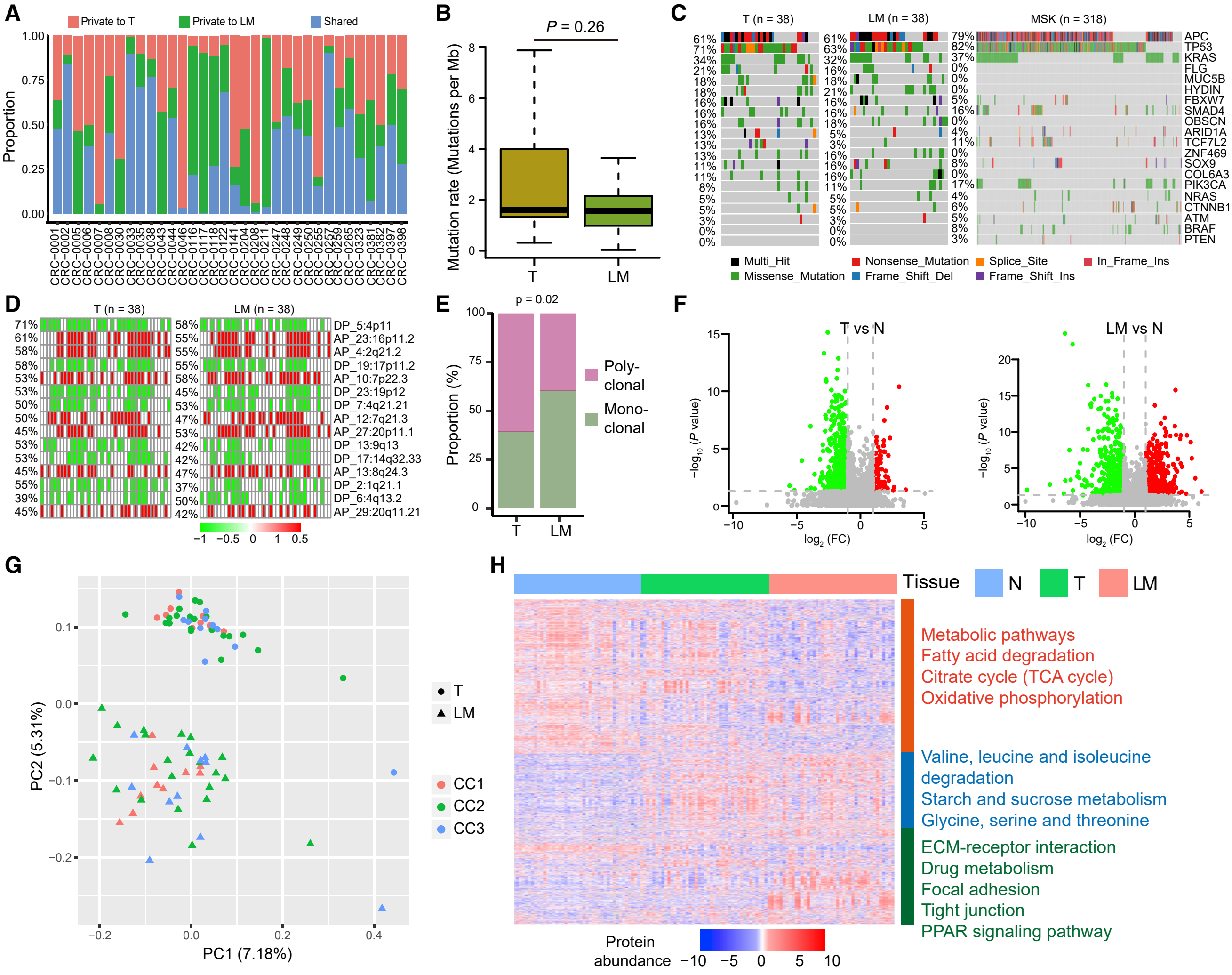


Figure 4. Proteogenomic Characteristics of mCRC

(A) The distribution of private and shared mutations in primary and metastatic tissues of mCRC.

(B) Comparison of mutation burdens between primary and metastatic mCRC tissues. The line and box represent median and upper and lower quartiles, respectively. The p value was calculated by two-tailed Student's t test.

(C) Comparison of frequently mutated genes in primary and metastatic tissues of mCRC in CCRC and metastatic tissues from MSK cohorts (Yaeger et al., 2018).

(D) Comparison of the top 15 frequent SCNAs in primary and metastatic tissues of mCRC.

(E) Clonal differences between primary and metastatic tissues of mCRC. The p value was calculated by chi-square test.

(F) Differentially expressed genes between primary (left) or metastatic (right) and remote normal tissues. Significance tests were performed by two-tailed Student's t test.

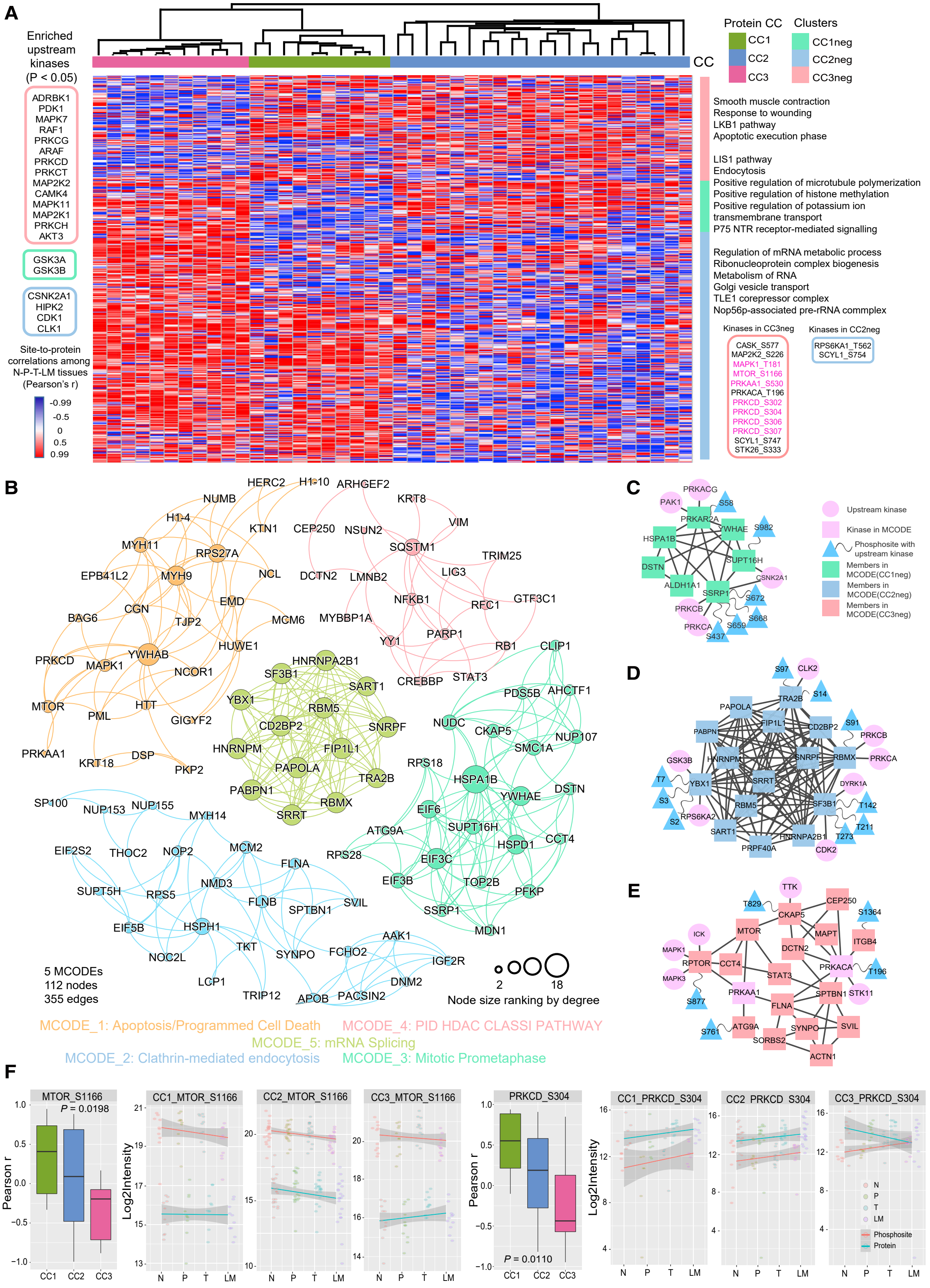
(G) Principal-component analysis plot of differentially expressed proteins indicated in (F).

(H) Normalized expression profiles of the differentially expressed proteins in N-T-LM tissues and their enriched pathways. N, remote normal tissue; T, primary tumor tissue; LM, liver metastatic tissue.

See also Figure S6 and Table S4.

the distribution of correlations was bimodal and clearly shifted to positive values (Figures S6D and S6E) in all CC subtypes.

To look for significantly co-regulated phosphosite-to-protein relationships among the three CC subtypes, ANOVA was used to successfully identify 954 pairwise phosphosite-to-protein correlations that were significantly different among the three CC subtypes (ANOVA, BH adjusted $p < 0.05$). We found that CC2 distinctly showed a greater number of negative regulatory interactions, while CC3 and CC1 showed more positive co-variations (Figure S6E). The 954 pairwise phosphosite-to-protein correlations could distinguish between the three proteomic subtypes for all 42 mCRC cases (Figure 5A; Table S5), and the 954 phosphosite-to-protein pairs could be classified into three clusters: CC1 negative (CC1neg), CC2 negative (CC2neg), or CC3 negative (CC3neg). Metascape analysis (Zhou et al., 2019) revealed that the proteins corresponding to phosphosite-to-protein pairs in the CC3neg cluster were enriched in smooth muscle contraction, response to wounding, LKB1 pathway, and apoptotic execution phase. CC1neg cluster members were preferentially enriched in LIS1 pathway, endocytosis, and



(legend on next page)

Please cite this article in press as: Li et al., Integrated Omics of Metastatic Colorectal Cancer, Cancer Cell (2020), https://doi.org/10.1016/j.ccell.2020.08.002

p75 NTR receptor-mediated signaling. Concurrently, CC1neg cluster members were significantly involved in regulation of mRNA metabolic process, ribonucleoprotein complex biogenesis, and Nop56p-associated pre-rRNA complex.

We further mapped the experimentally verified (Hornbeck et al., 2015) or predicted (Horn et al., 2014) kinase-substrate pairs to CC1neg, CC2neg, and CC3neg cluster members. Using hypergeometric distribution, we selected the enriched kinases for each cluster ($p < 0.05$, Figure 5A). Notably, CC3neg members were enriched with the maximum number of kinases, and, interestingly, the three clusters shared no common upstream kinases, indicating the contribution of a diversity of kinase-substrate networks among the three clusters. Among the 954 significant co-variations, 14 phosphosites corresponding to 10 kinases were significantly co-regulated (Figure 5A). Among them, 12 phosphosite-to-protein correlations, corresponding to 9 kinases (right pink box), belonged to the CC3neg cluster, while only 2 correlations were involved in the CC2neg cluster (right blue box). Notably, higher mutation frequencies were detected in 5 of these 10 kinases in CC3 (Tables S2 and S5), thus suggesting that the phosphosite-to-protein co-variation of kinases may be derived from genomic variation.

Using the MCODE complex/subnetwork analytical method (Bader and Hogue, 2003), we found five key co-varying phosphosite-to-protein MCODEs (hypergeometric test, $p < 0.001$), which included 112 nodes and 355 edges (Figure 5B; Table S5). Apoptosis, clathrin-mediated endocytosis, mitotic prometaphase, and HDAC class I pathway were the top four cooperative MCODEs (Figures S6F–S6I) in multiple tissues of mCRC. We noticed that transcriptional regulation by TP53 (Figure 5C) and the LKB1 pathway (Figure 5E) were the top CC1neg- or CC3neg-specific MCODEs, respectively. In contrast, mRNA splicing complexes (MCODE 5) was the top MCODE in CC2neg cluster members (Figure 5D). The experimentally verified (Hornbeck et al., 2015) or predicted (i.e., with highest NetworKIN score) (Horn et al., 2014) upstream kinases were also mapped. In the CC3neg cluster LKB1 pathway (Figure 5E), two kinases, PRKACA and PRKAA1, were also identified as members of MCODE. Although PRKCA and PRKCB were involved upstream of transcriptional regulation by TP53 (Figure 5C, CC1neg) and mRNA splicing (Figure 5D, CC2neg), most kinases were not shared within MCODEs of the three CCs.

Seven sites across four kinases, PRKCD, MAPK1, MTOR, and PRKAA1 (Figures 5A and S6F), were found in the CC3neg cluster in MCODE 1 apoptosis. The correlations between PRKCD-S304 or MTOR-S1166 and their corresponding proteins were distinct for each CC subtype, and they both showed differential protein or phosphosite expression profiles among the three CCs (Figure 5F). PRKCD-S304 was previously reported to be autophosphorylated and involved in many cancer types, and was also found to respond to temporal lapatinib suppression (Durgan et al., 2007; Imami et al., 2012). PKBalpha, a key regulator of cell growth, proliferation, and metabolism, was predicted to be the upstream kinase for MTOR-S1166. This phosphosite had also been observed to be upregulated during EGF stimulus and by EGF stimulus combined with MAPK inhibitors (Pan et al., 2009). Pathway enrichment analysis showed that the proteins overlapping (Table S5) between phosphosite-to-protein co-variation pairs and genes with high mutation frequency (Figure S6J), or differential SCNA genes in the three CC subtypes (Figure S6K), were predicted to participate in diverse pathways. Taken together, these findings indicated that the phosphosite-to-protein relationships across multiple tissues of a patient showed distinct characteristics associated with proteome subtypes.

Phosphoproteomics Profiling Provides Druggable Targets for mCRC

Considering that protein kinases have been developed as viable drug targets for cancer therapies (Knapp, 2018), we next inferred kinase activities based on differentially abundant phosphosites in the mCRC primary and metastatic tissues in each CC subtype by comparison with N. By performing kinase-substrate enrichment analysis (Wiredja et al., 2017), we found that different CCs were enriched for distinct kinases and that primary and metastatic tissues in the same CC showed different activities for the same kinase (Figure 6A; Table S6). CDK5 showed high activity in the primary tissue of CC1, but not the metastases of CC1 and other CCs (Figure 6A). Similarly, MAPK1 was highly enriched in both primary and metastatic tissues of CC3, but not in the other CCs (Figure 6A).

For kinases with quantifiable protein levels and clinically actionable drugs (Wu et al., 2019), we analyzed the corresponding phosphosubstrate abundances among the 42 paired N-T or N-LM tissues of mCRC (Figures 6B and 6C). In total, 251 pairs of kinase-phosphosubstrates were found by combining our quantitative proteomic data with PhosphoSitePlus (Hornbeck et al., 2015) or NetworKIN 3.0 (Horn et al., 2014) (Table S6). We observed high heterogeneity of phosphosubstrates and kinases in different tissues and proteomic subtypes of mCRC patients (Figure 6B). We next constructed kinase-substrate networks for primary and metastatic tumors in each proteomic subtype based on the Pearson correlation coefficient between each two pairs of kinase-phosphosubstrate (Figure 6C). Positive

Figure 5. Phosphosite-to-Protein Co-variation in Multiple mCRC Tissues

Pearson's correlation coefficients of four tissues (N, P, T, LM) between matched pairs of phosphosite abundances versus protein abundances were calculated. Hierarchical clustering analysis map of significantly changed phosphosite-to-protein correlations among three CC subtypes (ANOVA, BH adjusted $p < 0.05$). These phosphosite-to-protein pairs could be classed into (A) three clusters (CC1neg, CC2neg, and CC3neg) with differentially enriched functional annotations and (B–E) an interactome network. The enriched upstream kinases (hypergeometric test, BH adjusted $p < 0.05$) in three clusters are shown in the colored boxes on the left side of (A), while the significant phosphosite-to-protein correlations of kinases in the heatmap are given in the colored boxes on the right. (B) Top five phosphosite-to-protein co-variated MCODE complexes/subnetworks (hypergeometric test, $p < 0.001$). The most specific MCODEs in (C) CC1neg, (D) CC2neg, and (E) CC3neg were combined with upstream kinases. (F) Two examples, the Pearson's correlation coefficients between PRKCD-S304 and MTOR-S1166 abundances and the corresponding protein abundances, are shown for each CC on the left. Smooth lines for corresponding protein or phosphosite abundances among different tissues (Method = "lm, linear regression") and the related 95% confidence intervals (gray areas) are shown for each CC on the right. The line and box represent median and upper and lower quartiles, respectively. N, remote normal tissue; P, para-carcinoma tissue; T, primary tumor tissue; LM, liver metastatic tissue.
See also Figure S6 and Table S5.

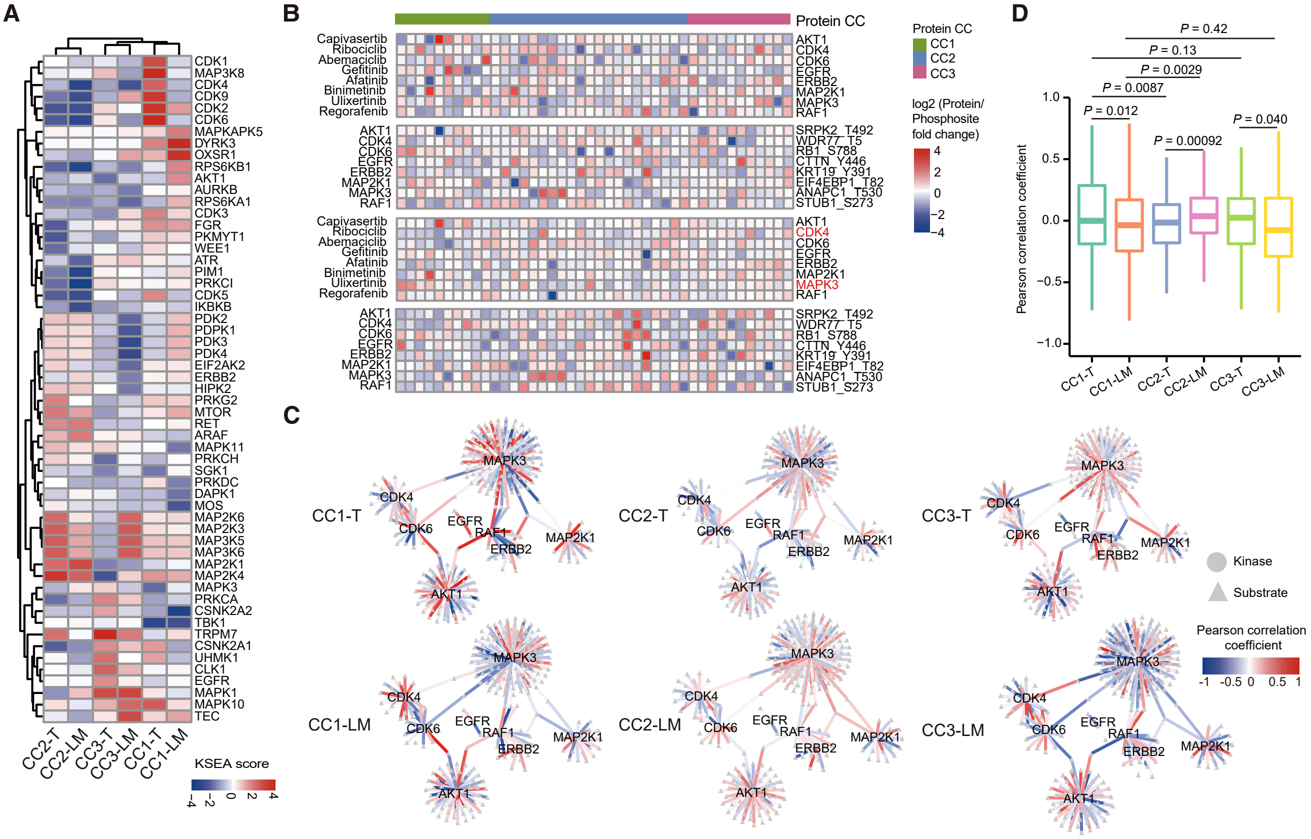


Figure 6. Phosphosubstrate Enrichment with Associated Kinases

(A) Kinase enrichment of differentially expressed phosphosites between T or LM and N in each CC.

(B) Phosphosubstrates of kinases with inhibitors and fold change at global and phospho levels for kinases and substrates, respectively. Names in red indicate kinases or phosphosubstrates demonstrating significant differences among the three proteomic subtypes. Significance tests were performed by ANOVA.

(C) Kinase-phosphosubstrate regulation networks in primary and metastatic tumors of each CC. The edges represent Pearson's correlation coefficient between kinases and the corresponding phosphosubstrates.

(D) Distribution of Pearson's correlation coefficients of kinase-substrate networks in (C). The p values were calculated with two-tailed Student's t test. The line and box represent median and upper and lower quartiles, respectively. T, primary tumor tissue; LM, liver metastatic tissue.

See also Figure S7 and Table S6.

correlations were observed between CDK4 and its substrates in CC1 metastatic tumors, whereas no significant correlations were found in CC1 primary tumors (Figure 6C). For MAPK3 and its phosphosubstrates, the largest differences were found between primary and metastatic tumors of the CC1 and CC3 subgroups (Figure 6C).

Significant differences were also found in the kinase-phosphosubstrate networks between primary and metastatic tumors in each proteomic subtype, and the CC3 networks showed greater similarity to CC1 than to CC2 networks (Figure 6D). The differences in kinase-phosphosubstrate networks between primary and metastatic tumors within subgroups were also larger than differences observed for primary tumors between subtypes. These observations suggest personalized and localization-specific responses to corresponding inhibitors in clinical treatments.

To further explore the potential for drug response in mCRC patients, we conducted pharmacological tests for three kinase inhibitors (afatinib, gefitinib, and regorafenib) on 31 miniPDX models, including nine pairs of primary-metastatic tumors and 13 other primary tumors (Zhang et al., 2018; Zhao et al., 2018) (Figure 7A). We measured the drug response effects of each tumor for each drug by tumor cell growth inhibition (TCGI, %) (Figure 7B; Table S7) and determined that primary and metastatic tumors from the same individual could exhibit different responses to the same drugs (Figure 7B). In particular, CCRC-0323 showed very high sensitivity to regorafenib (Figure 7B), which was explained by the mutations in the *RAF1* gene (Figure S7A). However, CCRC-0323 also harbored an *ERBB2* mutation, but showed no response to the *ERBB2* inhibitor afatinib (Figure 7B). By contrast, CCRC-0397 carried no *RAF1* mutation, but was sensitive to both inhibitors. In general, mutations were rarely found in genes corresponding to the three kinase inhibitors in the 31 miniPDX tumors (Figure S7A), thus excluding these patients from consideration for treatment with the corresponding drug in current clinical practice. However, our *in vivo* drug tests displayed good responses, suggesting that phosphoproteomic readout can be potentially more sensitive than the presence of mutations for indicating treatment suitability.

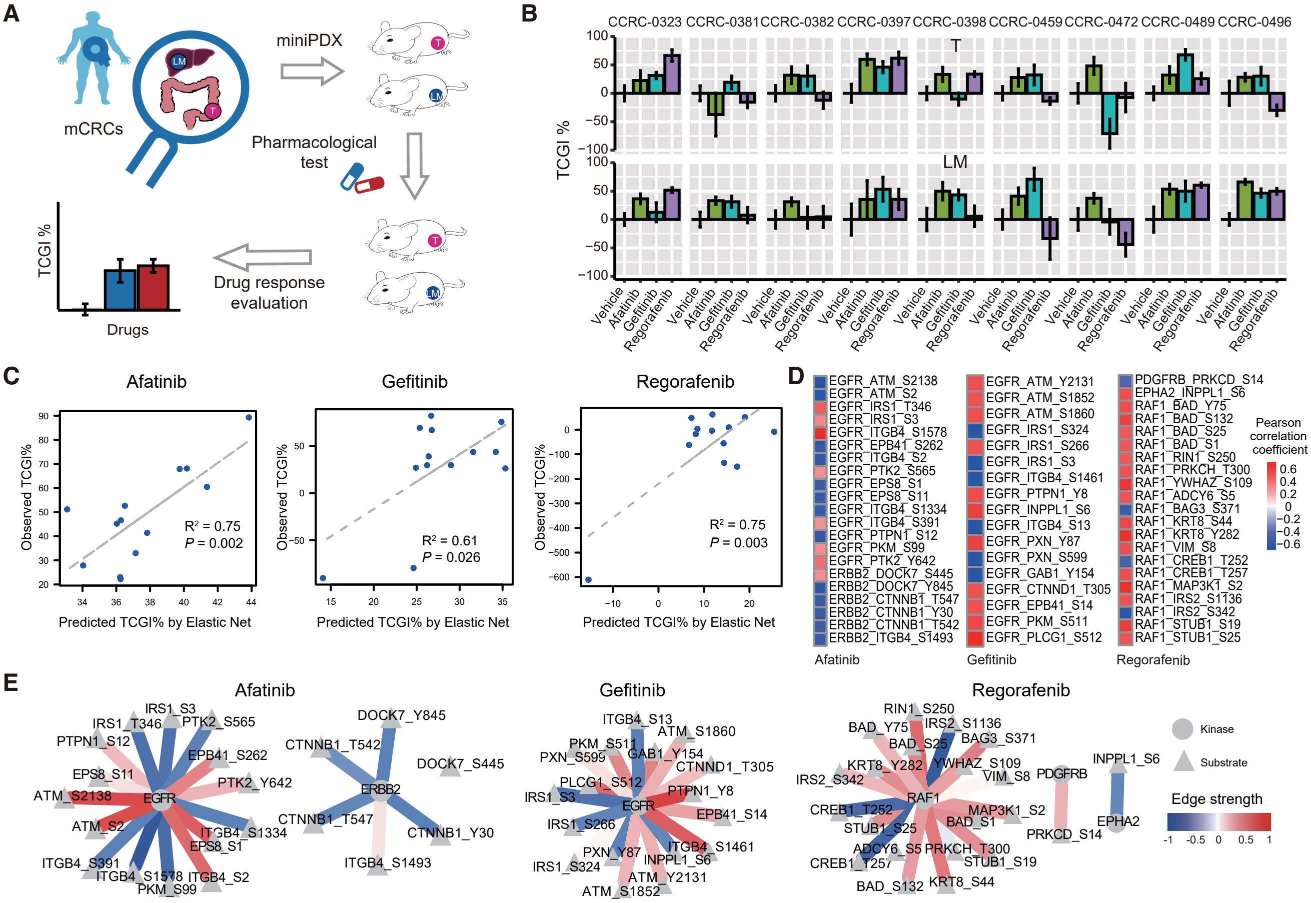


Figure 7. Kinase-Substrate Network Analysis and miniPDX Drug Tests

(A) Pharmacological tests using miniPDX models (Zhang et al., 2018; Zhao et al., 2018).

(B) Drug sensitivity results for primary and metastatic tissues of the 18 miniPDX models in training set. TCGI %, tumor cell growth inhibition. Mean ± SEM, n = 2 for each group.

(C) Correlations between elastic net-predicted and observed TCGI % in the validation set of 13 miniPDX models based on kinase-phosphosubstrate edge features. Correlations and p values were calculated by Pearson's correlation method.

(D) The Pearson's correlation coefficient between selected kinase-phosphosubstrate edge features and drug response (TCGI %) for afatinib, gefitinib, and regorafenib in training set.

(E) Kinase-phosphosubstrate networks of selected features in (D) for the three drugs. Correlations were calculated by Pearson's correlation coefficient. T, primary tumor tissue; LM, liver metastatic tissue.

See also Figure S7 and Table S7.

Discriminative Model to Guide Drug Selection

We next explored the correlations between kinase-phosphosubstrate networks and drug sensitivity. We first constructed tissue-specific kinase-substrate networks based on previously reported methods (Sun et al., 2019; Zhang et al., 2015a). Specifically, for each pair of kinase and substrate reported by PhosphoSitePlus (Hornbeck et al., 2015) or NetworKIN 3.0 (Horn et al., 2014), we calculated the edge strength between the kinase and the phosphosubstrate for both primary and metastatic tissues of each patient (Figure S7B). For each kinase and substrate pair, positive edge strength represents the same directional change between protein and phosphosite, while negative edge strength indicates anti-correlation (Figure S7B).

We then constructed elastic net regression models based on the 1,696 edge strength features for the prediction of drug responses (Table S7). Eighteen tumor tissues and corresponding drug test results were used as the training set, and the remaining 13 models were used as the validation set (Table S7). Remarkably, high correlations were observed between the predicted and the observed TCGIs for all three kinase inhibitors (Figure 7C). In total, 21, 17, and 21 pairs of kinase-phosphosubstrate edge features were selected by elastic net in the prediction of drug response to afatinib, gefitinib, and regorafenib, respectively (Figures 7D and 7E). Most of the selected features showed negative correlations with afatinib sensitivity, while positive correlations were observed between regorafenib sensitivity and the corresponding kinase-phosphosubstrate features (Figures 7D and 7E). Specifically, the positive correlation between EGFR and PTPN1-Y8 and the negative correlation between EGFR and IRS1-S3 both contributed to the afatinib treatment response (Figures 7D and 7E), thus suggesting that the response or sensitivity to treatment of tumors is not determined by a single factor nor exhibits strictly positive correlations, but rather reflects the sum of multiple molecular characteristics evident in multi-omics data. For regorafenib with multiple targets, the edges for RAF1, PDGRFB, and EHPA2 all contributed to drug response, while RAF1 contributed a predominant role in efficacy (Figures 7D and 7E). By contrast, models based on quantitative kinase and phosphosubstrate data showed poor accuracy in predicting drug response (Figures S7C and S7D). These results demonstrated the strong potential viability of kinase-phosphosubstrates networks for prediction of drug sensitivity of mCRC patients.

DISCUSSION

Here, we present a large-scale omics study on metastatic cancer, which demonstrates that proteomic patterns can distinctly classify primary tumors of CRC patients, with the power to discriminate potential prognostic outcomes. Particularly for stage IV patients, this classification may also significantly correlate with prognosis. Under current recommendations, stage I–III CRC can be treated with a variety of therapies that result in good 5-year survival rates. However, the progression and survival prediction for stage IV patients represents a highly challenging obstacle for successful treatment selection. The relationship between proteome pattern and prognosis can potentially facilitate the precise treatment and evaluation for stage IV patients.

Between the two groups of patients, i.e., with and without metastasis, the primary tumors harbored few differences in their mutational signatures. However, we found that combined proteomic and phosphoproteomic analysis provides the highest accuracy in distinguishing metastatic and non-metastatic patients, based on the patterns of data derived only from primary tumors. Moreover, proteogenomic analysis of primary tumor and concurrently metastatic tumor tissues revealed both similarities and variations between them. Our data also revealed an obvious loss of genomic mutations in metastatic tumors, which confirmed the recent results of a pan-cancer investigation (Priestley et al., 2019). However, we report here that the proteome and phosphoproteome of metastatic tissue from each individual can differ significantly from those of its primary counterpart.

Based on omics dataset and *in vivo* drug testing models, we established a machine learning model to predict drug response. For patients without druggable mutations, we proposed a strategy that exploits the protein-phosphorylation relationship to effectively select the most suitable targeted therapy. Promisingly, the accumulation of multi-omics data in conjunction with efficient drug testing can establish an accurate index for determining the most suitable drugs for a given tumor type.

In summary, our multi-omics data focused on mCRC provide the global characteristics of primary and metastatic tumors. Based on our observations of divergence in proteome and phosphoproteome patterns between metastatic and primary tumors, we found that network analysis combined with proteomics and phosphoproteomics data can accurately reflect drug responses. These results strongly suggest that the selection of therapies for stage IV patients should consider these proteome/phosphoproteome profiles in both primary and metastatic tissues for effective, individualized treatment strategies.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental Information can be found online at https://doi.org/10.1016/j.ccell.2020.08.002.

ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Programme of China (grant 2017YFA0505500), the Strategic Priority Research Program of the Chinese Academy of Sciences (CAS) (grants XDA12010000, XDB38000000), and the National Natural Science Foundation of China (NSFC) (grants 91029301, 30700397, 91529303, 31771476, 81572358). We thank Y.L.Y. and C.R.F. in the Big Data Platform (SIBS, CAS). We thank the staff in Shanghai LIDE Biotech for miniPDX modeling.

AUTHOR CONTRIBUTIONS

Conceptualization, W.J.R., Z.W., and Z.R.; Methodology & Investigation, L.C., S.Y.D., Y.G.Y., C.J.R., L.Z., Z.H., H.Y., B.C.G, D.L.L., L.P., Z.K., W.Y.H., and W.Q.Q.; Writing – Original Draft, L.C., S.Y.D., and Z.R.; Writing – Review & Editing, L.Q.R., W.Q.Q., L.Q., S.Y., L.Y.X., and C.L.N.; Supervision, W.J.R., Z.W., and Z.R.

DECLARATION OF INTERESTS

The authors declare no competing financial interests.

Received: January 8, 2020
Revised: June 22, 2020
Accepted: August 6, 2020
Published: September 3, 2020

REFERENCES

Alexandrov, L.B., Nik-Zainal, S., Wedge, D.C., Campbell, P.J., and Stratton, M.R. (2013). Deciphering signatures of mutational processes operative in human cancer. Cell Rep. *3*, 246–259.

Allison, K.H., and Sledge, G.W. (2014). Heterogeneity and cancer. Oncology *28*, 772–778.

Aryee, M.J., Jaffe, A.E., Corrada-Bravo, H., Ladd-Acosta, C., Feinberg, A.P., Hansen, K.D., and Irizarry, R.A. (2014). Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. Bioinformatics *30*, 1363–1369.

Austin, P.C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behav. Res. *46*, 399–424.

Bader, G.D., and Hogue, C.W. (2003). An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics *4*, 2.

Barkovits, K., Pacharra, S., Pfeiffer, K., Steinbach, S., Eisenacher, M., Marcus, K., and Uszkoreit, J. (2019). Reproducibility, specificity and accuracy of relative quantification using spectral library-based data-independent acquisition. Mol. Cell Proteomics *19*, 181–197.

Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., Kim, S., Wilson, C.J., Lehar, J., Kryukov, G.V., Sonkin, D., et al. (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature *483*, 603–607.

Bastian, M., Heymann, S., and Jacomy, M. (2009). Gephi: an open source software for exploring and manipulating networks. International AAAI Conference on Weblogs and Social Media.

Bhullar, K.S., Lagaron, N.O., McGowan, E.M., Parmar, I., Jha, A., Hubbard, B.P., and Rupasinghe, H.P.V. (2018). Kinase-targeted cancer therapies: progress, challenges and future directions. Mol. Cancer *17*, 48.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A., and Jemal, A. (2018). Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J. Clin. *68*, 394–424.

Bruderer, R., Bernhardt, O.M., Gandhi, T., Miladinovic, S.M., Cheng, L.Y., Messner, S., Ehrenberger, T., Zanotelli, V., Butscheid, Y., Escher, C., et al. (2015). Extending the limits of quantitative proteome profiling with data-independent acquisition and application to acetaminophen-treated three-dimensional liver microtissues. Mol. Cell Proteomics *14*, 1400–1410.

Cancer Genome Atlas, N. (2012). Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*, 330–337.

Cox, J., and Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. Nat. Biotechnol. *26*, 1367–1372.

Dekker, E., Tanis, P.J., Vleugels, J.L.A., Kasi, P.M., and Wallace, M.B. (2019). Colorectal cancer. Lancet *394*, 1467–1480.

Deng, Y. (2017). Rectal cancer in Asian vs. Western countries: why the variation in incidence? Curr. Treat. Opt. Oncol. *18*, 64.

Durgan, J., Michael, N., Totty, N., and Parker, P.J. (2007). Novel phosphorylation site markers of protein kinase C delta activation. FEBS Lett. *581*, 3377–3381.

Escher, C., Reiter, L., MacLean, B., Ossola, R., Herzog, F., Chilton, J., MacCoss, M.J., and Rinner, O. (2012). Using iRT, a normalized retention time for more targeted measurement of peptides. Proteomics *12*, 1111–1121.

Favero, F., Joshi, T., Marquard, A.M., Birkbak, N.J., Krzystanek, M., Li, Q., Szallasi, Z., and Eklund, A.C. (2015). Sequenza: allele-specific copy number and mutation profiles from tumor sequencing data. Ann. Oncol. *26*, 64–70.

Forbes, S.A., Beare, D., Boutselakis, H., Bamford, S., Bindal, N., Tate, J., Cole, C.G., Ward, S., Dawson, E., Ponting, L., et al. (2017). COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Res. *45*, D777–D783.

Gao, Q., Zhu, H., Dong, L., Shi, W., Chen, R., Song, Z., Huang, C., Li, J., Dong, X., Zhou, Y., et al. (2019). Integrated proteogenomic characterization of HBV-related hepatocellular carcinoma. Cell *179*, 1240.

Garnett, M.J., Edelman, E.J., Heidorn, S.J., Greenman, C.D., Dastur, A., Lau, K.W., Greninger, P., Thompson, I.R., Luo, X., Soares, J., et al. (2012). Systematic identification of genomic markers of drug sensitivity in cancer cells. Nature *483*, 570–575.

Gaujoux, R., Starosvetsky, E., Maimon, N., Vallania, F., Bar-Yoseph, H., Pressman, S., Weisshof, R., Goren, I., Rabinowitz, K., Waterman, M., et al. (2019). Cell-centred meta-analysis reveals baseline predictors of anti-TNFalpha non-response in biopsy and blood of patients with IBD. Gut *68*, 604–614.

Gerhauser, C., Favero, F., Risch, T., Simon, R., Feuerbach, L., Assenov, Y., Heckmann, D., Sidiropoulos, N., Waszak, S.M., Hubschmann, D., et al. (2018). Molecular evolution of early-onset prostate cancer identifies molecular risk markers and clinical trajectories. Cancer Cell *34*, 996–1011.e8.

Gilar, M., Olivova, P., Daly, A.E., and Gebler, J.C. (2005). Two-dimensional separation of peptides using RP-RP-HPLC system with different pH in first and second separation dimensions. J. Sep. Sci. *28*, 1694–1703.

Guinney, J., Dienstmann, R., Wang, X., de Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., et al. (2015). The consensus molecular subtypes of colorectal cancer. Nat. Med. *21*, 1350–1356.

Ho, H.Q., Honda, Y., Hamamoto, S., Ishii, T., Fujimoto, N., and Ishitsuka, E. (2018). Feasibility study of large-scale production of iodine-125 at the high temperature engineering test reactor. Appl. Radiat. Isot. *140*, 209–214.

Horn, H., Schoof, E.M., Kim, J., Robin, X., Miller, M.L., Diella, F., Palma, A., Cesareni, G., Jensen, L.J., and Linding, R. (2014). KinomeXplorer: an integrated platform for kinome biology studies. Nat. Methods *11*, 603–604.

Hornbeck, P.V., Zhang, B., Murray, B., Kornhauser, J.M., Latham, V., and Skrzypek, E. (2015). PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. Nucleic Acids Res. *43*, D512–D520.

Imami, K., Sugiyama, N., Imamura, H., Wakabayashi, M., Tomita, M., Taniguchi, M., Ueno, T., Toi, M., and Ishihama, Y. (2012). Temporal profiling of lapatinib-suppressed phosphorylation signals in EGFR/HER2 pathways. Mol. Cell Proteomics *11*, 1741–1757.

Imperiale, T.F., Abhyankar, P.R., Stump, T.E., and Emmett, T.W. (2018). Prevalence of advanced, precancerous colorectal neoplasms in black and white populations: a systematic review and meta-analysis. Gastroenterology *155*, 1776–1786.e1.

Knapp, S. (2018). New opportunities for kinase drug repurposing and target discovery. Br. J. Cancer *118*, 936–937.

Kroigard, A.B., Larsen, M.J., Laenkholm, A.V., Knoop, A.S., Jensen, J.D., Bak, M., Mollenhauer, J., Thomassen, M., and Kruse, T.A. (2018). Identification of metastasis driver genes by massive parallel sequencing of successive steps of breast cancer progression. PLoS One *13*, e0189887.

Kuipers, E.J., Grady, W.M., Lieberman, D., Seufferlein, T., Sung, J.J., Boelens, P.G., van de Velde, C.J., and Watanabe, T. (2015). Colorectal cancer. Nat. Rev. Dis. Primers *1*, 15065.

Lawrence, M.S., Stojanov, P., Polak, P., Kryukov, G.V., Cibulskis, K., Sivachenko, A., Carter, S.L., Stewart, C., Mermel, C.H., Roberts, S.A., et al. (2013). Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature *499*, 214–218.

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

McLaren, W., Gil, L., Hunt, S.E., Riat, H.S., Ritchie, G.R., Thormann, A., Flicek, P., and Cunningham, F. (2016). The ensembl variant effect predictor. Genome Biol. *17*, 122.

Mermel, C.H., Schumacher, S.E., Hill, B., Meyerson, M.L., Beroukhim, R., and Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. Genome Biol. *12*, R41.

Murphy, C.C., Wallace, K., Sandler, R.S., and Baron, J.A. (2019). Racial disparities in incidence of young-onset colorectal cancer and patient survival. Gastroenterology *156*, 958–965.

Myhre, S., Lingjaerde, O.C., Hennessy, B.T., Aure, M.R., Carey, M.S., Alsner, J., Tramm, T., Overgaard, J., Mills, G.B., Borresen-Dale, A.L., et al. (2013). Influence of DNA copy number and mRNA levels on the expression of breast cancer related proteins. Mol. Oncol. *7*, 704–718.

Naba, A., Clauser, K.R., Whittaker, C.A., Carr, S.A., Tanabe, K.K., and Hynes, R.O. (2014). Extracellular matrix signatures of human primary metastatic colon cancers and their metastases to liver. BMC Cancer *14*, 518.

Navarro, P., Kuharev, J., Gillet, L.C., Bernhardt, O.M., MacLean, B., Rost, H.L., Tate, S.A., Tsou, C.C., Reiter, L., Distler, U., et al. (2016). A multicenter study benchmarks software tools for label-free proteome quantification. Nat. Biotechnol. *34*, 1130–1136.

Niu, B., Ye, K., Zhang, Q., Lu, C., Xie, M., McLellan, M.D., Wendl, M.C., and Ding, L. (2014). MSIsensor: microsatellite instability detection using paired tumor-normal sequence data. Bioinformatics *30*, 1015–1016.

Pan, C., Olsen, J.V., Daub, H., and Mann, M. (2009). Global effects of kinase inhibitors on signaling networks revealed by quantitative phosphoproteomics. Mol. Cell Proteomics *8*, 2796–2808.

Priestley, P., Baber, J., Lolkema, M.P., Steeghs, N., de Bruijn, E., Shale, C., Duyvesteyn, K., Haidari, S., van Hoeck, A., Onstenk, W., et al. (2019). Pan-cancer whole-genome analyses of metastatic solid tumours. Nature *575*, 210–216.

Punt, C.J., Koopman, M., and Vermeulen, L. (2017). From tumour heterogeneity to advances in precision treatment of colorectal cancer. Nat. Rev. Clin. Oncol. *14*, 235–246.

Rappsilber, J., Mann, M., and Ishihama, Y. (2007). Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nat. Protoc. *2*, 1896–1906.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. *13*, 2498–2504.

Simon, M.S., Thomson, C.A., Pettijohn, E., Kato, I., Rodabough, R.J., Lane, D., Hubbell, F.A., O'Sullivan, M.J., Adams-Campbell, L., Mouton, C.P., et al. (2011). Racial differences in colorectal cancer incidence and mortality in the Women's Health Initiative. Cancer Epidemiol. Biomarkers Prev. *20*, 1368–1378.

Song, C., Ye, M., Han, G., Jiang, X., Wang, F., Yu, Z., Chen, R., and Zou, H. (2010). Reversed-phase-reversed-phase liquid chromatography approach with high orthogonality for multidimensional separation of phosphopeptides. Anal. Chem. *82*, 53–56.

Sun, Y., Li, C., Pang, S., Yao, Q., Chen, L., Li, Y., and Zeng, R. (2019). Kinase-substrate edge biomarkers provide a more accurate prognostic prediction in ER-negative breast cancer. Genomics Proteomics Bioinformatics.

Sung, J.J.Y., Chiu, H.M., Jung, K.W., Jun, J.K., Sekiguchi, M., Matsuda, T., and Kyaw, M.H. (2019). Increasing trend in young-onset colorectal cancer in asia: more cancers in men and more rectal cancers. Am. J. Gastroenterol. *114*, 322–329.

Tawk, R., Abner, A., Ashford, A., and Brown, C.P. (2015). Differences in colorectal cancer outcomes by race and insurance. Int. J. Environ. Res. Public Health *13*, https://doi.org/10.3390/ijerph13010048.

Thakur, S.S., Geiger, T., Chatterjee, B., Bandilla, P., Frohlich, F., Cox, J., and Mann, M. (2011). Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. Mol. Cell Proteomics *10*, M110 003699.

Tyanova, S., Temu, T., and Cox, J. (2016). The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat. Protoc. *11*, 2301–2319.

Vasaikar, S., Huang, C., Wang, X., Petyuk, V.A., Savage, S.R., Wen, B., Dou, Y., Zhang, Y., Shi, Z., Arshad, O.A., et al. (2019). Proteogenomic analysis of human colon cancer reveals new therapeutic opportunities. Cell *177*, 1035–1049.e9.

Wilkerson, M.D., and Hayes, D.N. (2010). ConsensusClusterPlus: a class discovery tool with confidence assessments and item tracking. Bioinformatics *26*, 1572–1573.

Wiredja, D.D., Koyuturk, M., and Chance, M.R. (2017). The KSEA App: a web-based tool for kinase activity inference from quantitative phosphoproteomics. Bioinformatics *33*, 3489–3491.

Wisniewski, J.R., Zougman, A., Nagaraj, N., and Mann, M. (2009). Universal sample preparation method for proteome analysis. Nat. Methods *6*, 359–362.

Wu, X., Xing, X., Dowlut, D., Zeng, Y., Liu, J., and Liu, X. (2019). Integrating phosphoproteomics into kinase-targeted cancer therapies in precision medicine. J. Proteomics *191*, 68–79.

Yaeger, R., Chatila, W.K., Lipsyc, M.D., Hechtman, J.F., Cercek, A., Sanchez-Vega, F., Jayakumaran, G., Middha, S., Zehir, A., Donoghue, M.T.A., et al. (2018). Clinical sequencing defines the genomic landscape of metastatic colorectal cancer. Cancer Cell *33*, 125–136.e3.

Zhang, B., Wang, J., Wang, X., Zhu, J., Liu, Q., Shi, Z., Chambers, M.C., Zimmerman, L.J., Shaddox, K.F., Kim, S., et al. (2014). Proteogenomic characterization of human colon and rectal cancer. Nature *513*, 382–387.

Zhang, F., Wang, W., Long, Y., Liu, H., Cheng, J., Guo, L., Li, R., Meng, C., Yu, S., Zhao, Q., et al. (2018). Characterization of drug responses of mini patient-derived xenografts in mice for predicting cancer patient clinical therapeutic response. Cancer Commun. (Lond) *38*, 60.

Zhang, W., Zeng, T., Liu, X., and Chen, L. (2015a). Diagnosing phenotypes of single-sample individuals by edge biomarkers. J. Mol. Cell Biol. *7*, 231–241.

Zhang, Y., Cai, Q., Shu, X.O., Gao, Y.T., Li, C., Zheng, W., and Long, J. (2015b). Whole-exome sequencing identifies novel somatic mutations in Chinese breast cancer patients. J. Mol. Genet. Med. *9*, 183.

Zhao, P., Chen, H., Wen, D., Mou, S., Zhang, F., and Zheng, S. (2018). Personalized treatment based on mini patient-derived xenografts and WES/RNA sequencing in a patient with metastatic duodenal adenocarcinoma. Cancer Commun. (Lond) *38*, 54.

Zhao, Y., and Jensen, O.N. (2009). Modification-specific proteomics: strategies for characterization of post-translational modifications using enrichment techniques. Proteomics *9*, 4632–4641.

Zhou, Y., Zhou, B., Pache, L., Chang, M., Khodabakhshi, A.H., Tanaseichuk, O., Benner, C., and Chanda, S.K. (2019). Metascape provides a biologist-oriented resource for the analysis of systems-level datasets. Nat. Commun. *10*, 1523.

Please cite this article in press as: Li et al., Integrated Omics of Metastatic Colorectal Cancer, Cancer Cell (2020), https://doi.org/10.1016/j.ccell.2020.08.002

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
N/A	N/A	N/A
Biological Samples		
Primary tumor tissues, liver metastatic tissues, para-carcinoma tissues, remote normal tissues and blood samples from Chinese patients with colorectal cancer (CCRC)	This study; ChangHai Hospital (Shanghai, China)	N/A
MiniPDX models	This study; ChangHai Hospital (Shanghai, China); Shanghai LIDE Biotech	N/A
Oligonucleotides		
N/A	N/A	N/A
Chemicals, Peptides, and Recombinant Proteins		
Afatinib	Tsbiochem	Cat# T1773; CAS:850140-73-7
Gefitinib	Bidepharm	Cat# BD131918; CAS: 184475-35-2
Regorafenib	Bidepharm	Cat# BD559192; CAS: 835621-07-3
Critical Commercial Assays		
QIAamp DNA mini kit	QIAGEN	Cat# 51306
SureSelect XT Human All Exon V6	Agilent	Cat# 5190-8864
BCA Protein Assay kit	Thermo Scientific	Cat# 23225
EZ DNA Methylation Kit	Zymo Research	Cat# D5002
Infinium MethylationEPIC BeadChip Kit	Illumina	Cat# 15073390
High-Select Fe-NTA kit	Thermo Scientific	Cat# A32992
Modified microencapsulation and hollow fiber culture system (OncoVee MiniPDX®)	Shanghai LIDE Biotech	N/A
Deposited Data		
Proteogenomic data of the CCRC cohort	This paper	NODE database: http://www.biosino.org/node/project/detail/OEP000729
COSMIC83; COSMIC - Catalogue Of Somatic Mutations In Cancer	(Forbes et al., 2017)	RRID: SCR_002260; https://cancer.sanger.ac.uk/cosmic
Proteogenomic characterization of human colon and rectal cancer	(Zhang et al., 2014)	https://cptac-data-portal.georgetown.edu/cptac/s/S022
Consensus Molecular Subtype (CMS) of colorectal carcinomas based on Gene Expression Profiles (GEP)	(Guinney et al., 2015)	https://www.nature.com/articles/nm.3967
CPTAC Cancer Proteome Confirmatory Colon Study	(Vasaikar et al., 2019)	https://cptac-data-portal.georgetown.edu/cptac/s/S037
MSK cohorts	(Yaeger et al., 2018)	https://www.cbioportal.org/study/summary?id=crc_msk_2017
PhosphoSitePlus	(Hornbeck et al., 2015)	RRID: SCR_001837; https://www.phosphosite.org/
NetworKIN 3.0	(Horn et al., 2014)	RRID: SCR_007818; http://networkin.info/
Experimental Models: Cell Lines		
293T	ATCC	Cat# ATCC® CRL-11268; RRID: CVCL_1926
CCD 841 CoN	ATCC	Cat# ATCC® CRL-1790; RRID: CVCL_2871
Caco2	ATCC	Cat# ATCC® HTB-37; RRID: CVCL_0025
Colo205	ATCC	Cat# ATCC® CCL-222; RRID: CVCL_0218
HCT116	ATCC	Cat# ATCC® CCL-247; RRID: CVCL_0291

(*Continued on next page*)

Please cite this article in press as: Li et al., Integrated Omics of Metastatic Colorectal Cancer, Cancer Cell (2020), https://doi.org/10.1016/j.ccell.2020.08.002

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
HT29	ATCC	Cat# ATCC® HTB-38; RRID: CVCL_0320
LOVO	ATCC	Cat# ATCC® CCL-229; RRID: CVCL_0399
SW620	ATCC	Cat# ATCC® CCL-227; RRID: CVCL_0547
Experimental Models: Organisms/Strains		
BALB/c-Foxn1nu/Nju	GemPharmatech	Cat# T000521
Oligonucleotides		
N/A	N/A	N/A
Recombinant DNA		
N/A	N/A	N/A
Software and Algorithms		
MaxQuant 1.6.2.10	(Cox and Mann, 2008)	RRID: SCR_014485; http://www.coxdocs.org/doku.php?id=maxquant:start
Spectronaut™ 13	Biognosys Inc.	https://www.biognosys.com/
Burrows-Wheeler Aligner (BWA) (version 0.7.15)	(Li and Durbin, 2009)	RRID: SCR_010910; http://bio-bwa.sourceforge.net/
Picard (version 2.5.0)	GitHub	RRID: SCR_006525; http://broadinstitute.github.io/picard/
Genome Analysis Toolkit (GATK) (version 4.0.11)	Broad Institute	RRID: SCR_001876; https://software.broadinstitute.org/gatk/
Mutect (version 2)	Broad Institute	https://software.broadinstitute.org/gatk/
Ensemble variant effect predictor (VEP v94.5)	(McLaren et al., 2016)	RRID: SCR_007931; https://asia.ensembl.org/info/docs/tools/vep/script/vep_download.html
MutSigCV (version 1.41)	(Lawrence et al., 2013)	https://software.broadinstitute.org/cancer/cga/mutsig
MSIsensor (v0.5)	(Niu et al., 2014)	RRID: SCR_006418; https://github.com/ding-lab/msisensor
Sequenza (version 3.0.0)	(Favero et al., 2015)	RRID: SCR_016662; https://cran.r-project.org/web/packages/sequenza/index.html
GISTIC (version 2.0.23)	(Mermel et al., 2011)	RRID: SCR_000151; https://portals.broadinstitute.org/cgi-bin/cancer/publications/view/216
MatchIt R package (version 3.0.2)	(Ho et al., 2018)	https://cran.r-project.org/web/packages/MatchIt/index.html
NMF R package (version 0.21.0)	(Gaujoux et al., 2018)	https://cran.r-project.org/web/packages/NMF/index.html
ConsensusClusterPlus R package (version 1.42.0)	(Wilkerson and Hayes, 2010)	RRID: SCR_016954; http://bioconductor.org/packages/release/bioc/html/ConsensusClusterPlus.html
CMSclassifier	(Guinney et al., 2015)	https://github.com/Sage-Bionetworks/CMSclassifier
KSEAapp (version 0.99.0)	(Wiredja et al., 2017)	https://cran.r-project.org/web/packages/KSEAapp/index.html
minfi	(Aryee et al., 2014)	RRID: SCR_012830; https://bioconductor.org/packages/release/bioc/html/minfi.html
Cytoscape	(Shannon et al., 2003)	RRID: SCR_003032; https://cytoscape.org/
Gephi	(Bastian et al., 2009)	RRID: SCR_004293; https://gephi.org/
R	R Core Team	RRID: SCR_002394; https://www.r-project.orgs
Other		
N/A	N/A	N/A

RESOURCE AVAILABILITY

Lead Contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Rong Zeng (zr@sibcb.ac.cn).

Materials Availability

The clinical materials, including established miniPDX models, from ChangHai Hospital (Shanghai, China) will have restrictions according to Institutional Review Board and Material Transfer Agreement institutional policies. Other materials not specified will be made available from the corresponding author on request.

Data and Code Availability

The raw data and processed datasets generated during this study are available at The National Omics Data Encyclopedia (NODE) under accession OEP000729 or through the link (http://www.biosino.org/node/project/detail/OEP000729). The processed mutation data, proteomic data, phosphoproteomic data and clinical data can be found in Table S1. Details for software availability are in the Key Resources Table.

EXPERIMENTAL MODEL AND SUBJECT DETAILS

Clinical Sample Acquisition

Chinese patients with CRC (CCRC) who took colorectal cancer surgery at Colorectal Surgery Department, Shanghai Changhai Hospital (Naval Medical University, Shanghai, China). We collected written informed consent from all participating patients during the establishment of the clinical sample bank. All these patients were followed with a median time of 1,240 days. Among them, 43 metastatic CRC (mCRC) patients also took their hepatectomy of liver metastatic cancer at same time. In addition to primary tumor tissues and liver metastatic tissues, we also gathered remote normal tissues (N, 5-cm-away from the tumor edge), para-carcinoma tissues (P, 2-cm-away from the tumor edge, normal adjacent tissues), and pre-operation blood samples as references. Each tissue specimen was collected within 30 min after resection, immediately transferred into sterile freezing vials and immersed in liquid nitrogen, cut into 0.5 cm^3 pieces under $-40°$C, then splited and stored at $-80°$C until use. As for blood samples, plasma and blood cell were sub-packaged into 500 μl per vial and stored at $-80°$C until use. The informed consent was obtained from all subjects. The experimental protocol was approved by Shanghai Changhai Hospital Ethics Committee (CHEC2017-235, Shanghai, China).

To established mini patient derived xenograft models (MiniPDX models), the study protocol of clinical samples was approved by the Institutional Ethics Committee of Shanghai LIDE Biotech. Tumor tissue acquisition was approved by the ethics committees of each participating hospital (Changhai Hospital, Shanghai, China) and agreed to by each patient via written informed consent and was carried out according to state and institutional regulations on experimental use of human tissues.

Cell Models

A total of 8 publicly available cell lines were purchased from American type culture collection (ATCC) by cell bank in Shanghai Institute of Biochemistry and Cell Biology. The 293T cell line was cultured in Dulbecco's Modified Eagle's Medium with 10% Fetal Bovine Serum (FBS); CCD 841 CoN was cultured in minimum Eagle's medium with 10% FBS; Caco2 was cultured in Eagle's Minimum Essential Medium with 20% FBS; Colo205 was cultured in RPMI-1640 Medium with 10% FBS; HCT116 and HT29 were cultured in McCoy's 5A Medium with 10% FBS; LOVO was cultured in ATCC-formulated F-12K Medium with 10% FBS; and SW620 was cultured in Leibovitz L-15 Medium with 10% FBS and 100% atmosphere. Identifiers for cell lines are in the Key Resources Table.

Animals

Five-week-old female nu/nu mice (BALB/c-Foxn1nu/Nju, T000521, GemPharmatech Co., Nanjing, China) were housed at the AAALAC accredited animal facility at LIDE Biotech (Shanghai, China). All animals received human care, and all study protocols were reviewed and approved by the Institutional Animal Care and Use Committee (IACUC) at LIDE Biotech, and conducted in accordance with established national and international regulations for laboratory animal protection.

METHOD DETAILS

Experimental Design

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment. We performed several quality-control steps to ensure the quality of samples used in the final cohort (Figure S2A). We selected the 146 CRC individuals for final analysis from 193 collected cases. We removed 21 cases without key baseline information, 15 cases with low protein identification due to low tumor purity (<50%), 3 cases with poor DNA quality and 8 cases without enough tissues for protein sample preparation.

Proteogenomic Workflow

The workflow of proteogenomic analysis for our CCRC cohort was shown in Figure S1. Accurate evaluation of tumor cellularity was determined by the middle section of each tumor tissue block, which was resected and subjected to hematoxylin and eosin (H&E) staining. The histological assessment of all tumor samples was accomplished by two board-certified pathologists independently (Chen-Guang Bai, Lu-Lu Deng). All the 146 primary tumor tissues (T) were confirmed to contain an average of 76% (Range 50–90%) tumor cell nuclei with averaged 12% necrosis (Range 2–40%). Meanwhile, 43 liver metastatic tissues (LM), 146 remote normal tissues (N, 5-cm-away from the tumor edge), 145 para-carcinoma tissues (P, 2-cm-away from the tumor edge, normal adjacent tissues), and 128 matched blood samples (BC, blood cell samples of one day before operation) were also included in the proteogenomic analysis. Detailed clinicopathologic features and prognosis information were summarized in Table S1, including gender, age, TNM stage, anatomic site, metastatic sites, survival or progression situations, and so on.

About 10∼25 mg wet-weight for each tissue sample or 150 μl for each blood sample was used for DNA extraction and whole exome sequencing (WES). Genomic DNA was extracted, then transported using dry ice to Berry Genomics (Beijing, China). About 20∼50 mg wet-weight tissue for each tissue sample was lysed with SDT buffer (4% w/v SDS, 100 mM Tris-HCl, 0.1 M DTT, pH 7.6) (Wisniewski et al., 2009) and stored at -80°C until use for the proteomic and phosphoproteomic analyses. Totally, 330 DNA samples (146 T, 38 LM, 128 BC, and 18 CN) and 480 protein samples (146 T, 43 LM, 146 CN, and 145 CP) were used in proteogenomic workflow (Figure S1, Table S1).

Whole Exome Sequencing

Genomic DNA were extracted from tumor or blood samples using the QIAamp DNA mini kit from QIAGEN. WES Library preparation was performed using SureSelect XT Human All Exon V6 kit (Agilent Technologies) according to manufacturer’s instructions. Pooled libraries were run on NovaSeq6000 (2x150 paired end runs) to achieve a minimum of 200x on target coverage for each sample library. The raw Illumina sequence data were demultiplexed and converted to fastq files subsequently. After adaptor removal and low-quality sequences trimming, qualified reads were used for somatic mutation detection, microsatellite instability prediction, and somatic copy number alteration (SCNA) analysis as described below.

Somatic Mutation Detection

For single nucleotide variant (SNV) and INDEL detection, we followed the Genome Analysis Toolkit (GATK, version 4.0.11) best practice guideline for somatic short variant discovery (https://software.broadinstitute.org/gatk/best-practices/workflow?id=11146). Briefly, we mapped qualified paired-end WES reads to human reference genome (hg19) with BWA-mem (v0.7.15-r1140). The BAM files were subsequently reordering, sorting, adding read groups and marking duplicates using picard tools (version 2.5.0). Then base quality score recalibration and INDEL realignment were performed using IndelRealigner and BaseRecalibrator functions of GATK. SNVs and INDELs (insertions/deletions) were called from tumor and matched-normal pairs using MuTect2 from GATK. Noted that blood cells were treated as normal for most of the 146 CCRC patients, and 18 remote normal tissues were used as the normal reference for 18 individuals with no available blood cell samples (Table S1). The sequence variants were then annotated using Ensemble variant effect predictor (VEP v94.5). The variants obtained by Ensemble VEP were filtered for protein altering events including non-synonymous SNVs, frameshift INDELs, non-frameshift INDELs, and stop gains. Significantly mutated genes were evaluated by MutSigCV (v1.41), and genes with a false discovery rate (q value) below 0.05 were considered significantly mutated above the background mutation rate.

Microsatellite Instability Prediction

To estimate microsatellite instability status for each patient, we applied MSIsensor (v0.5) (Niu et al., 2014) to compute length distributions of microsatellites per site in paired tumor and matched-normal BAM files, subsequently using these to statistically compare observed distributions in both samples. The total number of sites with sufficient data (at least 20 spanning reads in both normal and tumor) and also the number of somatic sites were calculated and the percentage of somatic sites is nominated as MSI score. Samples with MSI score >= 20% were assigned “MSI-H” as recommended by the algorithm.

Mutational Signature Analysis

Non-negative matrix factorization (NMF) approach (Alexandrov et al., 2013) was implemented to infer the mutation signatures jointly for the 146 primary tumors of the CCRC cohort. The 96 mutational contexts generated by somatic SNVs from six base substitutions (C > A, C > G, C > T, T > A, T > C, and T > G) within 16 possible combinations of neighboring bases for each substitution were used as input data to infer their contributions to observed mutations. The *de novo* detected mutation signatures were also compared to 30 known COSMIC cancer signatures (Alexandrov et al., 2013) to infer their exposure contributions.

Propensity Score Matching for Clinical Parameters

To eliminate the selection bias of patients in CCRC and TCGA databases caused by baseline demographic and clinical factors, propensity score matching (Austin, 2011) was used to balance the two datasets. Baseline clinicopathological characteristics including age at diagnosis, gender, AJCC stage and metastatic status were fit into a multivariate logistic model, and the nearest neighbor algorithm and one to one match were set in the logistic model. R package “MatchIt” was used for the calculation algorithm.

Somatic Copy Number Alteration (SCNA) Analysis

Somatic copy number alteration (SCNA) analysis used BAM files that were processed in the somatic mutation detection pipeline. These BAM files were further processed by R package sequenza (version 3.0.0) with default parameters (Favero et al., 2015). Sequenza calculates allele-specific copy number profiles based on a probabilistic model. Briefly, the python script 'sequenza-utils' firstly calculates the GC content in sliding windows from the genome reference hg19 and processes the sequencing data from the tumor and normal specimens for genomic positions with sufficient sequencing depth (>20 reads by default) to determine homozygous and heterozygous positions in the normal specimen and calculate the variant alleles and allelic frequency from the tumor specimen. Secondly, sequenza.extract efficiently performs GC-content normalization of the tumor versus normal depth ratio, and performs allele-specific segmentation using the 'copynumber' package. Third, sequenza.fit applies the Bayesian probabilistic model to infer cellularity and ploidy parameters and copy number profiles and the results were returned by sequenza.results. Gene-level SCNAs and significant SCNAs in the cohort were identified by Genomic Identification of Significant Targets in Cancer (GISTIC, version 2.0.23) to determine which SCNA regions were significantly gained or lost than expected by chance with q value % 0.1. We set up a threshold of 0.3 (-ta and -td parameters of GISTIC2) in picking the amplified or deleted regions based on the distribution of germline copy number variants. A log2 ratio cut-off of ± 0.3 was used to define SCNA amplification and deletion. We further summarize the arm-level copy number change based on a weighted sum approach (Vasaikar et al., 2019), in which the segment-level log2 copy ratios for all the segments located in the given arm were added up with the length of each segment being weighted.

Subclonal Copy-Number Analysis

Subclonal copy number were identified by computing the cancer cell fractions based on the B-allele frequency CCF_{bi} and the depth ratio CCF_{ri} for each segment by sequenza (Favero et al., 2015). A sample-wide analysis provides the copy number state estimate for each segment based on the total clonal contribution and the global cellularity (ρ) and ploidy (ψ) values.

Based on a previously described method (Gerhauser et al., 2018), we assume that subclones share the same ploidy but differ in cellularity. The CCF_{bi} and CCF_{ri} are calculated by dividing the estimated cellularity derived by the depth-ratio model ρ_{ri} and the B-allele frequency model ρ_{bi} with the sample-wide cellularity ρ: $CCF_{bi} = \rho_{bi}/\rho$ and $CCF_{ri} = \rho ri/\rho$. A bivariate Dirichlet process was then applied to generate 2D clusters of CCF_{bi} versus CCF_{ri}. Clusters with both CCF_{bi} and CCF_{ri} values between 0.1 and 0.9 were identified as subclonal clusters. Samples in which the sum of the subclonal segments represent more than 0.1% of the genome are classified as polyclonal, otherwise are classified as monoclonal.

DNA Methylation Data and Identification of MLH1 Hypermethylation

Based on mutation statues in MMR pathways and our proteomic subtyping, 32 primary tumors tissues were selected from our CCRC cohort for DNA methylation processing. About 10~25 mg wet-weight for each tissue sample was used for DNA extraction. Genomic DNA was extracted, then transported using dry ice to Shanghai Biotechnology Corporation (Shanghai, China). Genomic DNA from each sample was chemically modified and bisulfite-converted using the EZ DNA Methylation kit (Zymo Research Corp., Irvine, CA, USA) according to the manufacturer's instructions. According to user guide, Illumina Infinium® MethylationEPIC array BeadChip Kit (850K, 853,307 CpG sites, Illumina, San Diego, CA) was used to generate DNA methylation profiles. R package "minfi" was used to pre-process the raw array data and beta value (β) was used to represent methylation level for each probe site. Probes located within potential promoter regions (1500 bp flanking regions upstream and downstream of Transcription Start Sites (TSSs) of all transcripts annotated by UCSC) were examined. We next analyzed the *MLH1* hypermethylated samples across the 32 primary tumors. "Hypermethylated" sample at probe level was defined as the z-score normalized beta value > 1. A sample was identified as hypermethylated at the gene level if more than half the probes for *MLH1* were labeled as hypermethylated.

Protein Extraction and Digestion

Tissue samples or cell samples were minced, lysed in SDT lysis buffer followed by 5 min of heating at 95 °C and 3 min of sonication (5 second on and 10 second off, power 50 Watts). The lysate was clarified by centrifugation for 10 min at 14,000×g. Then, the supernatant was collected in new tube as whole extract and detected protein concentration using tryptophan-based fluorescence quantification method (Thakur et al., 2011). Protein sample was digested by filter-aided sample preparation protocol (FASP) (Wisniewski et al., 2009) using 10 kDa centrifugal filter tubes (Millipore). In 50 mM NH_4HCO_3 solution at 37 °C, trypsin (Promega) was added in two rounds. The first round was lasting 12 h with 1:50 of total protein amount, and the second round was lasting additional 4 h with equal trypsin amount. Each peptide mixture was eluted by centrifugation and dried by speed-vac.

Phosphopeptide Enrichment

The phosphopeptide enrichment was performed using High-Select Fe-NTA kit (Thermo Scientific) according to the kit manual and previous report (Gao et al., 2019) with some following modifications. In brief, the resins of one spin column in the kit were divided into 5 equal parts and mixed with each peptide mixture from 500 μg protein dissolved with 200 μL loading buffer (80% ACN, 0.1% TFA). The peptide-resin mixture was incubated for 30 min with thrice gentle blowing at room temperature, and then transferred into a home-packed one-layer Empore-C8 StageTip (Rappsilber et al., 2007) to remove nonspecific peptides and elute phosphopeptides. The elutes were immediately dried by speed-vac at 45 °C for mass spectrometry analysis.

Please cite this article in press as: Li et al., Integrated Omics of Metastatic Colorectal Cancer, Cancer Cell (2020), https://doi.org/10.1016/j.ccell.2020.08.002

High-pH RPLC Fractionation

As for spectral library for CCRC proteome or phosphoproteome, we used hybrid spectral library generation according to Spectronaut^{TM} 13 instructions (Biognosys Inc., released after June 2019) (Bruderer et al., 2015). One type of the spectral libraries was built by peptide fractionation in order to increase the depth of protein or phosphopeptide identification. Therefore, high-pH reverse phase liquid chromatography was used. Peptide mixtures were fractionated by a Waters XBridge BEH300 C18 column (250 x 4.6 mm, OD 5 mm) on Shimadzu Prominence HPLC System following manufacturer's instructions (Shimadzu Scientific Instruments). Mobile phases A and B were prepared based on previous paper (Gilar et al., 2005). As for proteome spectral library, a 97-min gradient was set as follows, 5%–7.5% B in 2 min; 7.5%–12% B in 5 min; 12%–25% B in 40 min; 25%–32% B in 25 min; 32%–95% B in 7 min; 95% B for 4 min; 95%–5% B in 4 min; 5% B for 10 min. The eluate was auto-collected every 1 min (Except to the last min) into 96 fractions. According to HPLC chromatogram, 30 fractions for proteome were combined by a concatenation scheme (Song et al., 2010). As for phosphoproteome spectral library, peptide mixtures digested from 30 mg proteins were used to fractionate (~3mg peptide per time, 5 times in total). A 85-min gradient was set as follows, 5%–7.5% B in 2 min; 7.5%–12% B in 5 min; 12%–25% B in 35 min; 25%–32% B in 22 min; 32%–95% B in 2 min; 95% B for 4 min; 95%–5% B in 4 min; 5% B for 11 min. The eluate was auto-collected every 2 min into 42 fractions. Next, according to HPLC chromatogram, 20 fractions were combined by a concatenation scheme (Song et al., 2010). And then, each fractionation was dried in a speed-vac and reconstituted to enrich phosphopeptide using High-Select Fe-NTA kit as described above. Each fraction was analyzed individually with LC-MS/MS settings as described below.

Benchmark for Nano-LC-MS/MS

The peptide retention time in the reverse phase chromatography could be converted into iRT space (Escher et al., 2012). If accurate iRTs are provided, the shotgun analysis will be speed up significantly, and the quality of results will be increase (Sensitivity, specificity, accuracy). To ensure calibration on difficult matrices and allow for detailed quality control readouts, we spiked equal amounts of iRT into each shotgun run not only in data-dependent acquisition (DDA) mode but also in data-independent acquisition (DIA) mode.

Each day for running DDA or DIA raw files, we also ran a shot of iRT alone to check chromatographic stability conveniently. The retention time for the first m/z in iRT peptides was ensured at about 4 ± 0.5 min, and all peak widths of iRT peptides was checked about 30 s by Xcalibur^{TM} software. As for each homemade column for running DDA or DIA raw files, we use accurately quantified 1μg 293T peptide mixture (BCA Protein Assay Kit, Thermo Scientific) to check the column pressure and column efficiency. At beginning, column pressure was about 150 bar in buffer A (0.1% formic acid). For each 293T raw, database searching against human UniProt database (Download on July 2017) was performed using MaxQuant 1.5.2.8 (Protein and peptide with false discovery rate < 0.01) (Tyanova et al., 2016). The average number of identified proteins, identified peptides, and average minute of retention length were 3202, 17821, and 0.396, respectively. Each standard sample was analyzed on a Thermo Scientific™ EASYnLC™ 1000 nanoflow LC. The sample was resolved using 0.1% formic acid and was separated using a home-made micro-tip C18 column (75 μm x 200 mm) packed with ReproSil-Pur C18-AQ, 3.0 mm resin (Dr. Maisch GmbH, Germany). Briefly, the sample was loaded onto a nano-C18 column and separated at a flow rate of 300 nL/min with following gradients: For iRT peptides, 0–1 min, 10% buffer B (0.1% fomic acid in acetonitrile); 1–13 min, 10–30% B; 13–15 min, 30–45% B; 15–16 min, 45–90% B; 16–22 min, 90% B. For 293T peptide mixture, 0–2 min, 5–8% buffer B; 2–42 min, 8–23% B; 42–50 min, 23–40% B; 50–52 min, 40–100% B; 52–60 min, 100% B. DDA performed 120K resolution MS scan and then triggered top 20 precursors (QE HF, Thermo Scientific Q Exactive HF hybrid quadrupole-Orbitrap mass spectrometer) for 15K resolution MS/MS scans. The MS or MS/MS AGC target value was set at 3e6 with 50 ms or 1e5 with 35 ms of max injection time, respectively.

Spectral Library

Hybrid library generation can offer a combined spectral library with both high depth (Project/sample-specific library) and high precision (DirectDIA library) (Barkovits et al., 2019; Navarro et al., 2016). Therefore, we utilized in-house generated sample-specific spectral libraries created using peptide fractionation approach, individual peptide samples and repeated DDA measurement. Meanwhile, high-quality and high-precision spectral libraries were also directly generated from multiple DIA data. In total, hybrid spectral library of CCRC proteome created by 165 DDA runs and 564 DIA runs, while hybrid spectral library of CCRC phosphoproteome created by 143 phospho-DDA runs and 519 phospho-DIA runs. Figure S2 was summarized the details of the two hybrid libraries. The hybrid spectral library of CCRC proteome included 179,382 precursors, 113,291 peptides, 11,510 protein groups and 9,942 gene products. The hybrid spectral library of CCRC phosphoproteome included 116,121 phospho-precursors, 65,851 phosphopeptides, 9,977 phosphoprotein groups and 7,125 phospho-gene products. Spectral library generated from the DDA files was searched with MaxQuant and built by Spectronaut^{TM}, and DirectDIA library generated by Spectronaut^{TM} as described below.

DDA and DIA Mode to Generate Proteomic or Phosphoproteomic Spectral Library

Equal amounts of iRT into each shotgun run of DDA and DIA mode. Each DDA or DIA sample was analyzed on a Thermo Scientific™ EASY-nLC™ 1000 nanoflow LC. The RP chromatographic column was the same as above. The sample was loaded onto a home-made nano-C18 column and separated at a flow rate of 300 nL/min with following gradients: For proteome in DDA and DIA mode, 0–2 min, 2–4% buffer B; 2–58 min, 4–30% B; 58–66 min, 30–45% B; 66–69 min, 45–90% B; 69–75 min, 90% B. For phosphoproteome in DDA and DIA mode, 0–2 min, 2–4% buffer B; 2–58 min, 4–23% B; 58–66 min, 23–40% B; 66–69 min, 40–90% B; 69–75 min, 90% B.

Data-dependent acquisition was performed using Xcalibur software in profile spectrum data type. A lock-mass m/z 445.12002 was used for internal calibration. The spray voltage was set at 2,300 V in positive ion mode and the ion transfer tube temperature was set at 270°C. DDA performed 120K resolution MS scan @ m/z 200 and then triggered top 20 precursors (QE HF). The MS AGC target value was set at 3e6 with 50 ms of max injection time by orbitrap mass analyzer (300-1,500 m/z). The MS/MS AGC target value was set at 1e5 with 35 ms of max injection time generated by HCD fragmentation (200-2,000 m/z) at a resolution of 15,000 @ m/z 200. The normalized collision energy (NCE) was set at NCE 28 %, and the dynamic exclusion time was 30 s. Precursors with charge 1, 7, 8 and > 8 were excluded for MS/MS analysis.

DDA files were processed using MaxQuant (1.6.2.10) with default settings. Carbamidomethyl (C) was set as fixed modifications. Oxidation (M), Acetyl (Protein N-term) were set as variable modifications. Reference FASTA files for human was downloaded from UniProt on July 2017, combining with the fusion sequence of iRT (Biognosys Inc.). In phosphorylation data analysis, phospho (STY) was also set as a variable modification. A maximum number of 5 modifications per peptide were allowed for each peptide. Enzyme specificity was set as trypsin/P. The maximum missing cleavage site was set as 2. The tolerances of first search and main search for peptides were set at 20 ppm and 4.5 ppm, respectively. The minimal peptide length was set at 7. The false discovery rates (FDR) of peptide, protein and site were all < 0.01.

Data-independent acquisition was also performed using Xcalibur software in profile spectrum data type. Basic parameters were equal to the DDA parameters described above. DIA isolation windows with variable width were decided by DDA searching results from MaxQuant. For proteome DIA MS runs, fragment analysis was subdivided into 27 DIA isolation windows of four different widths: 10 loop counts of 29 m/z with central m/z at 314.5, 343.5, 372.5, 401.5, 430.5, 459.5, 488.5, 517.5, 546.5, and 575.5; 11 loop counts of 28 m/z with central m/z at 604.0, 632.0, 660.0, 688.0, 716.0, 744.0, 772.0, 800.0, 828.0, 856.0, and 884.0; 5 loop counts of 55 m/z with central m/z at 925.5, 980.5, 1035.5, 1090.5, and 1145.5; 1 loop count of 300 m/z with central m/z at 1323.0. For phosphoproteome DIA MS runs, fragment analysis was subdivided into 34 DIA isolation windows of five different widths: 2 loop counts of 46 m/z with central m/z at 423.0 and 469.0; 4 loop counts of 24 m/z with central m/z at 504.0, 528.0, 552.0 and 576.0; 16 loop counts of 19 m/z with central m/z at 597.5, 616.5, 635.5, 654.5, 673.5, 692.5, 711.5, 730.5, 749.5, 768.5, 787.5, 806.5, 825.5, 844.5, 863.5 and 882.5; 10 loop counts of 21 m/z with central m/z at 902.5, 923.5, 944.5, 965.5, 986.5, 1007.5, 1028.5, 1049.5, 1070.5, and 1091.5; 2 loop counts of 99 m/z with central m/z at 1151.5 and 1250.5. MS scan was also performed before each DIA cycle.

Pulsar is Biognosys' proprietary search engine, integrated into Spectronaut^{TM} for library generation. Pulsar can search Thermo Scientific™ DDA and DIA data. Here, we used Pulsar to generate spectral libraries from both DDA (MaxQuant results) and DIA files with default settings. Human reference FASTA files was also downloaded from UniProt on July 2017, combining with the fusion sequence of iRT. False identifications were controlled by an FDR estimation (Cutoff 0.01) at three levels: peptide-spectrum match (PSM), peptide, and protein group level. For phosphoproteomic DIA runs, phospho (STY) was also set as a variable modification.

DIA Mode to Get Proteomic or Phosphoproteomic Data

As above, equal amount of iRT was mixed into each shotgun run of DIA mode, and nano-LC MS/MS method ran as same as described. In total, 480 proteome DIA runs and 480 phosphoproteome DIA runs (Figure S1) were analyzed by Spectronaut^{TM} against hybrid spectral library of CCRC proteome or phosphoproteome, respectively. Calibration was set to non-linear iRT calibration with precision iRT enabled. Identification was performed using 5% q-value cutoff on precursor and protein level while the maximum number of decoys was set to a fraction of 0.1 of library size. Quantity was determined on MS/MS level using area of XIC peaks with enabled cross run normalization. For phosphoproteomic analysis, minor quantified (Peptide) grouping was set by modified sequence and PTM localization was activated and probability cutoff set to 0, in order to summarize phosphopeptide or phosphosite later. Phosphosite quantification was counted from the quantity of phosphorylation sequences by Perl script according to MaxQuant strategy. As a result, we present the first proteome and phosphoproteome of metastatic CRCs, which included 8,450 quantified protein groups and 47,786 quantified phosphosites. In order to evaluate our DIA technology, proteome and phosphoproteome samples from 7 cell lines, 1 Homo sapiens colon normal cell (CCD 841 CoN) and 6 CRC cells (Caco2, Colo205, HCT116, HT29, LOVO, and SW620), were also gathered for duplicated DIA data. Figure S2 were summarized the details of data in total and among different sample groups.

Consensus Clustering for Proteomic and Phosphosproteomic Data

For proteomic subtype prediction, we applied consensus clustering on 2440 differentially expressed proteins between 146 primary tumors and paired normal tissues using ConseususClusterPlus R package with the following parameters: maxK = 10, reps = 1,000 bootstraps, pItem = 0.8, pFeature = 1, clusterAlg = "kmdist", distance = "spearman". The number of clustering was determined by three factors, the average pairwise consensus matrix within consensus clusters, the delta plot of the relative change in the area under the cumulative distribution function (CDF) curve, and the average silhouette distance for consensus clusters. We selected a 3-cluster as the best solution for the consensus matrix with k = 3 deemed to be a cleanest separation among clusters (Figure S6). We next applied predefined signature genes to the protein expression matrix to assign prospective tumors to previously identified proteomic subtypes (ProS A-E) from Zhang et al. (Zhang et al., 2014). We also employed the random forest predictor implemented in the R package CMSclassifier (https://github.com/Sage-Bionetworks/CMSclassifier) (Guinney et al., 2015) to assign the consensus molecular subtypes (CMSs) to each sample based on the protein expression matrix. Using a default posterior probability of 0.5 as a threshold for sample classification, we assigned our 146 prospective tumors to the four CMS subtypes (Guinney et al., 2015). For phosphoproteomic subtyping, the same parameter was used for the 1487 differentially expressed phosphosites compared with related N tissues, k = 2 was selected for each phosphoproteomic subtype (Figure 3).

Survival Analysis

Survival curves were generated using the Kaplan–Meier method, and the log-rank test was applied to calculate differences between the curves. Hazard ratios (HRs) and their 95% confidence intervals (CI) were estimated for each multivariate survival analysis using Cox proportional hazards models by the R package ‘‘survival’’.

Differential Mutations, SCNAs, Proteins and Phosphosites identification

To find mutations with differential mutational rate between mCRC and non-mCRC, or among the three proteomic subtypes, Chi’ square tests were performed and P values less than 0.05 were considered as significant. For SCNAs, proteins and phosphosites, we performed two-sided Student’s t tests to identify significantly differenced features between primary and remote normal tissues, or between mCRC and non-mCRC primary tissues. ANOVA test was used to identify differential SCNAs and proteins among three proteomic subtypes or among four tissues of mCRC patients. Differential phosphosites among six phosphoproteomic subtypes were identified by ANOVA test. Benjamini-Hochberg corrected P values less than 0.05 were considered as significant.

Phosphosite-to-protein Co-variation Analysis

As for 42 mCRC cases with all four tissue types (N, P, T, and LM), we merged their protein and phosphosite data by UniProt ID. The log2 transformed median normalized protein or phosphosite abundances were used for the following analysis. In each case, we computed the Pearson’s correlation coefficients of all four tissues between each matched pair of phosphosite abundances versus protein abundances using cor.test function in R. In total, we obtained an array of correlation coefficients with 13362 rows and 42 columns, corresponding to 13362 pairwise phosphosite-to-protein relationships and 42 mCRC cases. For given phosphosite-to-protein pair in each mCRC case, if the Pearson’s correlation coefficient exactly equal plus 1 or minus 1, the correlation values were removed. Which mean that only if at least 3 tissues have the protein and phosphosite abundances at the same time, the correlation values will be remained for further calculation. Among the 42 mCRC cases, 10, 21, and 11 cases were belonging to three consensus clusters (Figure 2) CC1, CC2, and CC3, respectively. We got 327,586 correlation values corresponding to 13362 phosphosite-to-protein pairs, where 75,089, 151,584, and 100,912 values were belonging to CC1, CC2, and CC3, respectively. Taken CC1, CC2, and CC3 cases as three groups, 954 significantly co-regulated phosphosite-to-protein correlations (ANOVA test, BH adjusted P value < 0.05) were selected to do HCA plot and further annotations. Totally, 25,729 correlation values were corresponding to the 954 significantly co-regulated phosphosite-to-protein pairs, where 6,016, 11,963, and 7,750 values were belonging to CC1, CC2, and CC3, respectively. Using density plot, we drew out the distribution of correlations in all and in significantly co-regulated phosphosite-to-protein pairs (Figure S6).

The significantly co-regulated phosphosite-to-protein pairs could be classed into 3 clusters (CC1neg, CC2neg, and CC3neg). Metascape analysis (Zhou et al., 2019) was used to do functional enrichment, and interactome analysis for the corresponding proteins of the three HCA clusters by default analysis parameters. Enriched functional terms required to include $\geq$ 3 candidates, Hypergeometric test P$\leq$ 0.01, and enrichment factor $\geq$ 1.5. Top 20 significantly enriched pathways for corresponding proteins of 954 significant phosphosite-to-protein pairs were selected (Hypergeometric test, multi-test-corrected q-values, Table S5). Interactome analysis to networks limited containing 3 to 500 candidate proteins using BioGrid, InWeb_IM, and OmniPath databases. Top 5 MCODE complexes extracted from the interactome network formed by proteins based on the combination of the corresponding protein lists in CC1neg, CC2neg, and CC3neg clusters. The nodes in Figures S6F–S6I and 5 represented the protein components of given MCODE complex, where each color encoded its origin. We used Gephi (https://gephi.org/) to visualize the top 5 MCODE complexes in Figure 5B. In each HCA cluster, only top 1 MCODE complexes were also shown in Figure 5 and Table S5. What’s more, MCODE network components were assigned biological ‘‘meanings’’, where top three best P value terms were retained (Hypergeometric test P < 0.001, Table S5). Functional enrichment analysis in Figures S6J–S6K were also done by Metascape analysis with default parameters.

For each HCA cluster, known or predicted up-stream kinases of the corresponding phosphosites in the significantly co-regulated phosphosite-to-protein correlations were provided based on PhosphoSitePlus® (PSP) (Hornbeck et al., 2015) or NetworKIN 3.0 (NetworKIN Score > 1) (Horn et al., 2014). The enriched up-stream kinases were selected by Hypergeometric test of relative substrate numbers in each cluster (BH adjusted P value < 0.05, Figure 5A, Table S5).

Kinase Activity Prediction

To estimate changes in a kinase’s activity, we performed kinase enrichment analysis on significantly differentiated phosphosites in tumor compared to matched normal for each subtype by kinase-substrate enrichment analysis (KSEA) (Wiredja et al., 2017). Known kinase-substrate site relationships from PhosphoSitePlus® (PSP) (Hornbeck et al., 2015) or NetworKIN 3.0 (Horn et al., 2014) with score more than 1 were used as the K-S sources. A kinase score was given for each kinase based exclusively on the collective phosphorylation status of its substrates and transformed into z-score. For the kinase enrichment analysis, the threshold for significantly enriched kinases was Benjamini-Hochberg corrected FDR less than 0.05.

Network Analysis

For subgroup based networks, Pearson correlation coefficient was calculated for each pair of kinase and substrate from PhosphoSitePlus® (PSP) (Hornbeck et al., 2015) or NetworKIN 3.0 (Horn et al., 2014). For single sample based networks, kinase

Please cite this article in press as: Li et al., Integrated Omics of Metastatic Colorectal Cancer, Cancer Cell (2020), https://doi.org/10.1016/j.ccell.2020.08.002

and phosphor-substrate edge features were constructed based on the correlation between each kinase and phosphor-substrate pair according to the method previously described (Zhang et al., 2015a). The transformation is described below.

$$\begin{matrix} kinase,\ u \\ phospho-substrate,\ v \end{matrix} \begin{pmatrix} x_{u,j,k} \\ x_{v,j,k} \end{pmatrix} \quad -> edge{<}u-v{>}_k \left(\frac{x_{u,j,k}-\mu_{u,k}}{\sigma_{u,k}} . \frac{x_{v,j,k}-\mu_{v,k}}{\sigma_{v,k}} \right)$$

where $x_{u,j,k}$ represents the original value of uth kinase in jth sample from kth class, $x_{v,j,k}$ represents the original value of vth phospho-substrate in jth sample from kth class, and k was set to 1 or 2, representing the T or LM tissue, respectively. In addition, $\mu_{u,k} = \frac{1}{n_k}\sum_{j=1}^{n_k}(x_{u,j,k} - \mu_{u,k})$ and $\mu_{v,k} = \frac{1}{n_k}\sum_{j=1}^{n_k}(x_{v,j,k} - \mu_{v,k})$ are sample means of kinase u and phosphor-substrate v, and $\sigma_{u,k} = \sqrt{\frac{1}{n_k}\sum_{j=1}^{n_k}(x_{u,j,k} - \mu_{u,k})^2}$ and $\sigma_{v,k} = \sqrt{\frac{1}{n_k}\sum_{j=1}^{n_k}(x_{v,j,k} - \mu_{v,k})^2}$ are the corresponding uncorrected sample standard deviation.

In Vivo Drug Response Test

To rapidly test drug efficacy *in vivo*, we established mini patient derived xenograft models (MiniPDX models) (Zhang et al., 2018; Zhao et al., 2018) for 9 pairs of T-LM tissues in 9 CCRC cases and 13 non-pairs of T tissues in other 13 CCRC cases (Table S7) according to the previous papers. In brief, fresh surgical tumor specimens were acquired from mCRC patients at Changhai Hospital. Tumor samples were washed with Hank's balanced salt solution (HBSS) to remove mucus and necrotic tumor tissues. After morselization, the tumor tissues were digested with collagenase at 37 °C for 1–2 h. Tumor cells were pelleted by centrifugation at 600g for 5 min followed by removal of blood cells and fibroblasts with magnetic beads. Cells were then washed with HBSS, and then filled into OncoVee® capsules (LIDE Biotech, Shanghai, China). Each capsule contained about 2,000 cells. Capsules were implanted subcutaneously via a small skin incision with 3 capsules per mouse (5-week-old female nu/nu mouse). Mice bearing MiniPDX capsules were treated with appropriate control or drugs (Vehicle control, Afatinib, Gefitinib, or Regorafenib). Afatinib, Gefitinib, and Regorafenib were all administered orally, as single administration (Daily [qd]×1) for continuous 7 days with a dose of 20 mg/kg, 75 mg/kg, or 30 mg/kg body weight, respectively. All these drugs were dissolved by pre-prepared in 0.5% hydroxypropyl methylcellulose (HPMC) and 0.2% Tween-80 solution. Vehicle controls were isometric 0.5% HPMC and 0.2% Tween-80 solution and correlated mice were treated as same way. Each treatment (Control or drugs) was done in sextuplicate capsules. After all capsules were removed from mice, tumor cell proliferation in each capsule was measured using the CellTiter Glo Luminescent Cell Viability Assay kit (G7571, Promega, Madison, WI, US). Tumor cell growth inhibition (TCGI) (%) was calculated using the published formula (Zhang et al., 2018).

Drug Sensitivity Prediction Model

To find drug response associated features and build prediction models, the 31 miniPDX models were split into two datasets. The paired 18 models were used as the training dataset, and the other 13 independent models were regarded as testing dataset (Table S7). Elastic net (EN) algorithm is powerful to create parsimonious models from a large number of features and a relatively small number of samples, and has been successfully used to build drug sensitivity prediction models in several studies (Barretina et al., 2012; Garnett et al., 2012). We firstly performed a pre-selection step for the kinase and phospho-substrate features based on their Pearson's correlation coefficients with examined drug sensitivity in the training set, then Elastic net regression models were built for each drug based on the selected kinase and phosphor-substrate node or edge features. These models were used to predict the drug response of the 13 miniPDX models in the testing cohort. Pearson correlation coefficients were calculated between the predicted drug sensitivity and examined ones to assess the prediction performance.

QUANTIFICATION AND STATISTICAL ANALYSIS

Statistical details of experiments and analyses can be found in the figure legends and main text above. Statistical significance tests, including Fisher's exact test, Chi-square test, Student's t-test, Anova test, and Pearson or Spearman correlation test were performed using R, as denoted in each analysis. Data in the barplot are presented as mean ± SEM (technical or biological replicates from miniPDX model). For box-and-whisker plot, the box indicates interquartile range (IQR), the line in the box indicates the median, the whiskers indicate points within Q3+1.5⊆IQR and Q1–1.5⊆IQR. Q1 and Q3, the first and third quartiles, respectively. All statistical tests were two-sided, and statistical significance was considered when P value < 0.05. To account for multiple-testing, the P values were adjusted using the Benjamini-Hochberg FDR correction. Kaplan–Meier plots (Log-rank test) were used to describe relapse-free survival. Significant factors were further subjected to a multivariate Cox regression analysis.