Multi-omic Analysis of Familial Adenomatous Polyposis Reveals Molecular Pathways and Polyclonal Spreading Associated with Early Tumorigenesis

Aaron M. Horning
Stanford University https://orcid.org/0000-0003-3247-0798
Edward D. Esplin
Invitae https://orcid.org/0000-0001-9205-3756
Si Wu
Stanford University
Casey Hanson
Stanford University
Nasim Bararpour
Stanford University
Stephanie A. Nevins
Stanford University
Lihua Jiang
Stanford University
Kévin Contrepois
Stanford University
Hayan Lee
Stanford University https://orcid.org/0000-0003-0571-3192
Tuhin K. Guha
Stanford University
Zheng Hu
Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences https://orcid.org/0000-0003-1552-0060
Rozelle Laquindanum
Stanford University
Meredith A. Mills
Stanford University
Hassan Chaib
Stanford University
Roxanne Chiu

Stanford University
Ruiqi Jian
Stanford University
Joanne Chan
Stanford University
Mathew Ellenberger
Seer Bio https://orcid.org/0000-0002-0965-9528
Winston R. Becker
Stanford University https://orcid.org/0000-0001-7876-5060
Bahareh Bahmani
Stanford University
Aziz Khan
Stanford University https://orcid.org/0000-0002-6459-6224
Basil Michael
Stanford University
D. Glen Esplin
Animal Reference Pathology
Jeanne Shen
Stanford University https://orcid.org/0000-0002-1519-0308
Samuel Lancaster
Stanford University
Uri Ladabaum
Stanford University
Anshul Kundaje
Stanford University https://orcid.org/0000-0003-3084-2287
Teri A. Longacre
Stanford University
William J. Greenleaf
Stanford University https://orcid.org/0000-0003-1409-3095
Christina Curtis
Stanford University School of Medicine https://orcid.org/0000-0003-0166-3802
James M. Ford
Stanford University
Michael P. Snyder (mpsnyder@stanford.edu)
Stanford University School of Medicine https://orcid.org/0000-0003-0784-7987

Article

Keywords: Familial adenomatous polyposis, APC gene, colorectal cancer, cancer evolution, multi-omic integration, pre-cancer, methylome, transcriptomics, proteomics, metabolomics

Posted Date: May 19th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-515393/v1

Abstract

Familial adenomatous polyposis (FAP) is a genetic disease causing hundreds of premalignant polyps in affected patients, leading to colorectal cancer (CRC), and is an ideal model to study early transition to CRC. We performed deep multi-omic profiling of 135 normal mucosal, benign and dysplastic polyps and adenocarcinoma samples from 6 FAP patients who consented to broad data sharing. Whole genome sequencing indicates that spatially separated polyps from the same donor harbor numerous mutations in common, but evolve independently, consistent with a model of polyclonal origin and spreading. Transcriptomic, proteomic, metabolomic and lipidomic analyses revealed a dynamic choreography of thousands of molecular and cellular events that occur during early hyperplasia, dysplasia and cancer formation. These involve processes such as cell proliferation, immune response, alterations in metabolism (including amino acids, lipids), hormones, and extracellular matrix proteins. Interestingly, activation of the arachidonic acid pathway was found to occur early in hyperplasia; this pathway is targeted by aspirin/NSAIDs, a common preventative treatment of FAP patients. Overall, our results reveal key genomic, cellular and molecular events during the earliest steps in cancer formation and potential mechanisms of pharmaceutical prophylaxis.

Introduction

Colorectal adenocarcinoma (CRC) is the third leading cause of cancer death in the United States, with more than 140,000 new cases each year, and is thus a major public health burden[1]. Understanding the clonal evolution and molecular pathways involved in the genesis of both sporadic and hereditary CRCs is critical for improved diagnosis, as well as for developing preventive and therapeutic interventions[2]. To date, most large-scale genomic studies, including The Cancer Genome Atlas (TCGA) and the Pan-cancer Analysis of Whole Genomes (PCAWG) have focused on sporadic CRCs [3,4]. These studies have uncovered genes and signalling pathways altered in invasive CRCs and corroborate the Vogelgram multi-hit model of CRC development, mutations in APC initiate the process of colon carcinogenesis, followed by mutations in genes such as *TP53, KRAS* and *SMAD4* [5–9]. However, the timing of somatic genomic alterations that arise prior to invasive transformation cannot readily be inferred from the genomic profiles of established tumors. Moreover, the earliest molecular and biochemical alterations involved in CRC tumorigenesis remain poorly understood as precursor lesions have been minimally studied as has the surrounding mucosa and dysplastic tissue.

Familial adenomatous polyposis (FAP) is an ideal system in which to study early events during CRC formation for multiple reasons. First, more than 95% of FAP patients harbour germline mutations in APC, a component of the Wnt signaling pathway which is somatically altered in 93% of all CRCs and presumed to be the initiating event [3,7]. Second, FAP patients develop hundreds to thousands of polyps in the colon, often starting in adolescence where small (<0.5 mm) polyps appear benign and those >0.5 mm to 1 cm usually exhibit dysplasia. Third, FAP patients have a nearly 100% lifetime risk of developing invasive adenocarcinomas, and thus often have their entire colon removed as prophylactic treatment once the

number of polyps that develop becomes unmanageable through endoscopic surveillance. Accordingly, molecular analysis of multiple polyps from the same patient, including those that are benign or dysplastic, as well as “normal mucosa” and in some cases, paired adenocarcinomas, can reveal the mechanisms by which APC mutations initiate premalignant polyps, and their transition to a malignant state, across a common germline genetic background. Such information is expected to aid our understanding of the requisite alterations for transformation in both hereditary and sporadic CRC and may inform the approach to prevention, early detection and treatment [10,11].

To better understand events that occur during colorectal hyperplasia, dysplasia and cancer formation, and as part of the Human Tumor Atlas Network (HTAN) [12], we performed whole-genome sequencing as well as transcriptome, methylome, proteome, metabolome and lipidome analyses of 135 samples (6-43 per patient) representing different stages of cancer progression - unaffected colonic mucosa, benign polyps, dysplastic polyps and adenocarcinoma from 6 FAP patients who consented to broad data sharing. These data reveal extensive subclonal spreading of early arising mutations across distant regions of the colon. Further, they delineate cellular and biochemical changes during early tumorigenesis through the transition to invasive disease, providing insights into mechanisms of CRC progression and its prevention. This rich open-access multi-omic, multi-sample dataset also provides a valuable scientific resource to the community, complementing the existing datasets on sporadic CRCs.

Results

Multi-omic analysis of FAP samples

We collected and analyzed 135 samples from the colons of six FAP patients (whose characteristics are shown in Supplemental Table 1), isolated from multiple regions (ascending, transverse, descending including sigmoid, and rectum) (Fig 1A). Samples spanned a range of sizes and degrees of dysplasia as determined using histological staining: 29 polyp-adjacent normal mucosa (M, hereafter referred to in the text as "mucosa"), 35 benign polyps (usually small) (B), 57 dysplastic polyps (small to large) (D), eight samples from three adenocarcinomas (AdCa), and six patient-matched blood samples. The percent of tumor and degree of dysplasia in the dysplastic polyps and adenocarcinoma was scored by a trained pathologist (see Methods).

Samples were subject to a battery of multi-omic assays including whole-genome and/or exome sequencing (WGS/WES), bulk tissue RNA-seq, proteomics (TMT labeling), untargeted metabolomics, and targeted lipidomics (Fig 1C). Additionally, eight samples were profiled via whole genome bisulfite sequencing (methylomics). High-resolution “single-cell” assays, including single nuclei (sn) RNAseq, snATACseq) are described in an accompanying manuscript[13]. Importantly, all donors in this study consented to an “open consent”, facilitating broad sharing of this multi-sample, multi-omic dataset.

The landscape of somatic alterations in familial adenomatous polyposis

Whole genome sequencing (WGS) was performed on germline and somatic DNA isolated from 108 samples (i.e. buffy coat (germline), mucosa, benign and dysplastic polyps, and adenocarcinomas) from 6 FAP patients. The tissue and germline samples were sequenced to either 30X or 60X average depth coverage and the adenocarcinoma samples were sequenced to 100X coverage (Supplemental Figure 1). Somatic single nucleotide variants (SNVs) and small insertion/deletions (Indels) were detected using Mutect2 in multi-sample mode (DNA isolated from blood was used as a germline reference) to increase sensitivity of detection of identical mutations occuring in different tissue samples within the same patient. Copy number variants (CNVs) as well as purity and ploidy estimates for each of the mucosa, polyp, and adenocarcinomas (AdCa) samples were detected using Titan [14] . We found 2203 ± 1726, 4863 ± 1707, 4083 ± 1985, 172,453 ± 203,555 (mean ± sd) SNVs and 1497 ± 1182, 3089 ± 825, 2036 ± 715, 49032 ± 51384 (mean ± sd) indels in the mucosa, benign, dysplastic and AdCa samples, respectively. Interestingly, stage accounts for most of the variance associated with the number of SNVs detected when compared with size of the lesion and dysplasia status (linear mixed effect model, p = 0.0005, Supplemental Figure 1B).

Across the 106 samples, a total of 302 non-synonymous and deleterious (CADD scores ≥ 20) SNVs, indels and CNVs were found in CRC-specific driver genes (Fig 2A). Although most mutations were found in the AdCa samples (from patients A001, F and G), there were numerous CRC driver mutations which appeared in the early benign polyps as well. We identified patient-specific germline APC mutations in each of their tissue samples and manually recovered them in 2 samples, 1 sample in patient F and G each, which had low variant sequencing depth in the affected APC region. An additional patient, A014, did not have an APC mutation detectable by WGS or clinical diagnostic genetic testing; consistent with the observation that 5% of FAP patients lack detectable germline APC mutations [15]. Somatic APC mutation status is associated with higher tumor stage (Chi-squared test; X = 21.01, p-value 1e-04), and somatic *KRAS* mutation exhibits a similar relationship (Chi-squared test; X = 16.01, p-value 0.00113) (Fig 2B). Although secondary APC and KRAS mutations may not be sufficient for polyp development, they are common across polyps with somatic *APC*, *KRAS* and *TP53* mutations are present in 43%, 28%, and 11% samples, respectively. This frequency is similar to the rates observed in other analyses of FAP samples [16].

In total, 63 genes across all samples were inferred to be under “positive selection” as determined using the ratio of nonsynonymous relative to synonymous mutations (dN/dS ratio) (Supplemental Figure 2K, FDR < 0.05) [17]. Many of these genes are involved in the immunoregulatory gene set (R-HSA-198933) from Reactome [18], including *FCGR3A, HLA-A, HLA-C, KIR2DL3, KIR2DL4, LILRA2, KRAS, APC, GNAS,* and *OTOP1* (Log(q-value) = -2.273, hypergeometric distribution) (Supplemental Table 3, Supplemental Figure 2K). Furthermore, immune-related mutations appear to be under positive selection early in polyp development as *LILRA2* is positively selected (dN/dS >1) in the benign and dysplastic samples (q=4.11e-6, q=2.5e-6, respectively), whereas *FCGR3A* (q = 1.04e−02) and *HLA-A* (q = 8.61e−04) exhibited evidence of positively selected in the dysplastic samples (Supplemental Figure 3K). Pairwise comparisons of positively selected genes (dN/dS > 1) by clinical stage indicates enrichment of *LILRA2* mutations in the benign stage and that *KRAS* and *APC* mutations are enriched in dysplastic polyps and adenocarcinomas (pairwise Fisher's

exact test, FDR < 0.05, Supplemental Figure 2L). This is consistent with immune dysregulation occurring early in polyp development and persisting in established AdCa, similar to the patterns seen at the single cell level in the accompanying manuscript (Becker et al, submitted).

CNV analysis indicates that alterations in chromosomes 14, 7 and 20, which are commonly gained in CRC, can be acquired in the benign and dysplasia stages of polyp development (Figure 2C) [19]. A positive association between the size of the lesion and the number of CNVs was observed (R^2 = 0.51, p = 2e-16) (Supplemental Figure 2A,B,C). Whereas mucosa samples are usually unaffected by amplifications and losses, dysplastic polyps have significantly more genomic changes when compared to benign polyps (p = 8.2e-6) (Supplemental Figure 2D,E,F). Genomic losses were prominent in adenocarcinomas, possibly indicating that this is an important step in transition to invasion (Figure 2C, Supplemental Figures 2D-G). Moreover, linear discriminant analysis predicted dysplasia and polyp stage from the fraction of the genome affected by CNVs, (Supplemental Figure 2J; 78% accuracy).

Polyp development and clonal evolution

Analysis of somatic mutations from multiple samples of the same patient provides insights into evolution of hyperplastic polyps located at spatially distinct regions in the colon. To characterize these developmental relationships, we analyzed patterns of somatic mutations across tissues from the same patient. We sought to investigate these patterns using multiple approaches. First, we examined the genetic relationship amongst samples from individual patients, by constructing phylogenetic trees based on mutations and small insertion/deletions (SNVs and InDels) based on the maximum parsimony method (Figure 3A, Supplemental Figures 3B,E,H). Deleterious cancer driver genes (CADD>=20) and pathogenic (CLINSIG) mutations are indicated on the phylogenetic tree (Figure 3B, Supplemental Figures 3A,F,I,L,O) [20–22]. Importantly, tree shape reflects the evolutionary histories of multiple lesions from a given patient, as illustrated for Patient A001. Here, the majority of samples harbor many common variants accrued over relatively short branch lengths, suggesting a shared developmental phase, prior to an extended independent growth phase, resulting in thousands of private mutations unique to branches corresponding to individual lesions. Indeed, identical mutations, including canonical drivers such as KRAS p.G12V mutations, were found in benign polyps along the transverse and ascending colon and in the adenocarcinoma in the descending colon, reflecting dramatic physical separation throughout the colon. The somatic origin of these shared events was confirmed by manually inspecting the supporting reads (Supplemental Figures 3R,S,T,U,V,W,X).

To further investigate the phylogenetic relationship between samples harboring identical mutations, we performed ancestral sequence reconstruction (ASR) to determine the likely locations of the mutations on the phylogenetic trees: private mutations on terminal branches and shared mutations on internal branches. We estimated the cancer cell fraction (CCF) for each mutation, thus accounting for differences in sample purity and ploidy[23]. Mutations present in mucosal tissues as well as many benign and dysplastic samples tended to have low CCF values, reflective of multiple subclonal populations that had not fixed in the population (Figure 3D, Supplemental Figure 3C,G,J,M,P). Most private mutations were

subclonal, reflecting abundant clonal heterogeneity, consistent with other studies in adenomas [24]. Notably, the distribution of CCF values correlates with tissue stage (linear discriminant analysis), indicating that lower heterogeneity is associated with advanced stage (Supplemental Figure 2J), consistent with the notion that selectively advantageous mutations fix in the population during progression to increasingly malignant states, contributing to reduced heterogeneity. We performed the same analyses on the five other FAP patients with similar results for each case; tissue lesions, including polyps, from different regions of the colon shared many mutations, but also acquired numerous private mutations (Supplemental Figure 1A, Supplemental Figures 3A,C,F,G,I,J,L,M,O,P), indicative of an early shared lineage followed by a long period of independent growth. As a control, we compared mutations across polyps from different individuals and observed few shared mutations linking different patients, as expected (Supplemental Figure 4M).

To determine the relative timing of deleterious SNV/Indel mutation events (colon cancer driver mutations with CADD>=20 and “pathogenic” CLINSIG annotations), we used the phylogenetic relationships between the mutations (shared or private) as well as clonal status to place the mutations from all 6 of the patients on a relative timeline of events (Figure 3E). The rationale behind the relative timeline of events is as follows: 1) because “shared” mutations are detected in more samples, they likely arise before private mutations and 2) because clonal mutations are more abundant than subclonal ones, they likely arise earlier. Using a Bradley Terry model, we find that benign and dysplastic polyps often haror *KRAS* mutations followed by *FBXW7* then second-hit *APC* mutations. This mirrors prior reports in sporadic CRCs [9] where *APC* mutations were inferred to occur early, followed by *KRAS* and *FBXW7* mutations and subsequent chr 5q deletion (spanning the *APC* gene). On a per patient level, we find evidence for mutations in *HLA-B* in the benign, dysplastic and adenocarcinoma samples (Patient F), potentially reflective of immune evasion as a result of disabled antigen presentation (Figure 3F). Additionally, we find *KMT2C* mutations in mucosa, dysplastic and adenocarcinoma samples (Patients A001 and G, Figures 3B,C,D,E, Supplementary Figures 3P,Q), suggesting a potential early role for epigenetic dysregulation in colon cancer development.

Quantification of subclonal sharing

WGS facilitated the assessment of subclonal sharing across lesions representing different disease stages and physical locations within the colon. In particular, we compared CCF values for every sample in a pairwise manner, revealing, an abundance of shared subclonal mutations between samples despite the comparatively high number of unique mutations (Figure 4A shows a representative example for A001, Supplemental Figures 4B,D,F,H,J). We quantified this relationship by calculating the Jaccard similarity index (JSI), in which the number of identical subclonal mutations is divided by the sum of private clonal mutations and shared subclonal mutations, as previously described[25] . High JSI values occur when two samples share a large number of subclonal mutations with few private clonal mutations, whereas low JSI values occur when two samples have fewer shared subclonal mutations and more private clonal mutations. Pairwise comparisons indicate extensive subclonal sharing across all samples but this varies with disease stage (Supplemental Figure 4L). For example, JSI values were extremely high (0.78 ± 0.02)

when comparing mucosa samples across all regions of the colon (Figure 4B,C and Supplemental Figures 4A,C,E,G,I,K). Evaluating pairwise comparisons across stages of malignant progression reveals a reduction in JSI values at later stages (Figure 4B,C and Supplemental Figures 4A,C,E,G,I,K). This is consistent with the fixation of selectively advantageous mutations in malignant lesions and is reflected in the correlation between number of total private mutations per sample and stage (Supplemental Figure 1B, $p<0.0001$, F-value = 17.2, linear mixed effects model). As might be expected, we find that as the number of lesions a subclonal mutation is detected in increases, its median cellular abundance (CCF) increases (Supplemental Figure 4N).

Additionally, for four patients, A001, A002, A014 and A015, we recorded detailed tissue sample location within the colon, enabling comparisons of the physical distance measured in centimeters with genetic similarity based on the JSI. Irrespective of stage, physical tissue lesion distance did not correlate with genetic similarity, again suggesting that shared subclonal mutations are developmentally early events (Figure 4B, C).

We further sought to investigate whether the patterns of genetic variation across polyps were compatible with a model of monoclonal or polyclonal origin. Briefly, we simulated polyp growth assuming a stochastic birth-death process beginning from a single embryonic progenitor cell in the early colonic epithelium, whereby polyps are initiated from either a single (monoclonal) or group of progenitor cells (n=20 or 40) within this field and grow under defined parameters (Methods). The observed patterns of genetic variation across polyps were remarkably consistent with those simulated under a model of polyclonal origin (either 20-40 cells) (Supplemental Figure 4O-V). These results are in line with an early case study by Thirwell et al. [26] who demonstrated the polyclonal nature of adenomas via clonal marking.

Collectively, the phylogenetics and ancestral sharing analyses as well as the patterns of subclonal mutation sharing support a model of hereditary cancer formation and clonal spreading in which mutations give rise to distinct clones early in development, possibly within the endoderm during gastrulation (Figure 4D). These clones disperse throughout the maturing colon as a result of mechanical forces and tissue expansion and continuously accrue mutations during cell division, resulting in polyclonal polyps.

Extensive molecular changes during polyp formation and progression

In order to better understand the molecular steps and pathways activated during the early stages of polyp formation and cancer progression, we performed deep multi-omic profiling of histologically normal mucosa (M), benign polyps (B), dysplastic polyps (D) and a limited number of adenocarcinomas (A) (Figure 1D). RNA-Sequencing, TMT-based proteomics, untargeted metabolomics and targeted lipidomics were performed on 114, 89, 77 and 77 samples, respectively. Because of limited material, particularly for the small polyps, not all assays could be performed for all samples. The relative levels of transcripts (n = 35,081), proteins (n = 12,389), metabolites (n = 1,157) and lipids (n = 514) were obtained after data curation and normalization (see methods).

PCA analysis of each omic layer revealed a continuous transition from mucosa to benign polyps to dysplastic polyps and for RNA-Seq, to adenocarcinoma, suggesting a gradual progression of malignancy (Fig 5A). Analysis of differences relative to normal mucosa revealed thousands of molecular changes. The largest number of changes were found for transcripts and the least for metabolites and lipids (Fig 5B, Supplemental Data 3, 4, 5 and 6). At the gene expression level (Supplemental Data 3), most changes between cancer samples were evident in transitions to adenocarcinoma, with many of these same changes occurring at the early stages (i.e. from mucosa to benign and to dysplasia) (Fig. 5C). Similarly, differential analysis of proteomics, metabolomics and lipidomics (Supplemental Data 4, 5 and 6) revealed that the most significantly changed proteins, metabolites and lipids occur at the early stages (Fig. 5C; see methods & Supplemental Figure 5).

For the three early stage normal mucosa to benign polyp (M-B), normal mucosa to dysplasia (M-D), and benign polyp to dysplasia (B-D) and the late transitions to AdCa (M-A, B-A, D-A), transcripts and proteins evincing significant changes in abundance were subject to pathway enrichment via Qiagen Ingenuity Pathway Analysis (IPA). Similar analyses were performed for metabolite and lipid species using Metabolite Set Enrichment Analysis (MSEA) [27], ConsensusPathDB's [28]over-representation analysis and topology based pathway analysis [29] (see methods). These analyses revealed a total of 842 unique pathways that were altered (Summarized in Fig 6A and Fig 6B; molecules and pathways are summarized in Supplemental Data 7 and 8 respectively). We discuss the molecular features associated with these transitions in the following sections.

Transition from mucosa to benign polyp

To identify the earliest events involved in polyp formation, we examined the molecules and pathways altered in benign polyps relative to FAP mucosa and found 4,017 molecules and over one hundred pathways (Fig. 5B, Fig. 5C and Fig. 6; Supplemental Data 8; corrected p-value < 0.05). A number of connected pathways are shown in Fig. 7A. Notably, the arachidonic acid pathway was upregulated in benign polyps relative to FAP mucosa as indicated by key metabolites and proteins (metabolomics; FDR=1E-03 /lipidomics;FDR=7E-03) which also had alterations in the levels of key proteins (e.g. the *GPX* family (*GPX 2,3,4*; FDR=4.6E-05), *PTGES3* (FDR=2.9E-02), *PTGDS1* (FDR=3.1E-03), *PTGIS* (FDR=2.8E-03), *TBXAS1*(FDR=4.46E-03), *DPEP1*(FDR=1.72E-03), *LTA4H* (FDR= 2.1E-02) (Fig. 7B). The arachidonic acid pathway is particularly interesting since it is involved in inflammation and can be suppressed by aspirin/NSAIDs, a standard treatment for FAP patients to slow polyp progression and reduce cancer formation [30]. Thus, our results provide a molecular explanation for the therapeutic response of FAP-treated patients.

We also observed an extensive alteration of immune response pathways at the transcript and protein levels. Among 82 dysregulated immune response pathways, 26 were supported by both RNA-seq and proteomics data (e.g. complement system, IL-12 signaling and production in macrophages, production of nitric oxide and reactive oxygen species in macrophages, etc), and the remaining pathways (n = 56) were only significant at the transcript level (e.g. agranulocyte adhesion and diapedesis, Th1 and Th2

activation pathway, phagosome formation, T cell exhaustion signaling pathway, etc). The complete list of immune response pathways and their changes in each transition are shown in Supplemental Figure 6. We also detected several immune response-related pathways at metabolite and lipid levels (e.g. Ca-dependent event enrichment -log10 FDR = 5.0E-04). Altogether, these observations may indicate a predominant role of the immune system in the early progression of colon cancer at many different molecular and cellular levels (summarized in Fig 7C).

Additional pathways involved in cell proliferation and cell cycle (Transcriptome, Figure. 6A) and amino acid and lipid metabolism (Metabolome/lipidome Figure 6B) were altered in benign polyps. In particular, we identified a depletion of TAG stores in the benign polyps (Supplemental Figure 7), which was supported by alterations in transcripts levels of several key TAG biosynthesis genes, such as GPAT3, AGPAT family, PLPP family, DGAT1/2 (Fig. 7A). Lower abundance of TAGs may suggest utilization of this lipid class as a key source of energy supply for polyp growth. Significant depletion of TAGs has been reported previously in cancerous tissue [31,32]; our results suggest that this is an early event during polyp formation.

Although in many cases pathway enrichment analysis using the different assays (e.g. transcriptome and proteome) resulted in similar results, some changes were only evident with a single assay. For instance, the cell cycle/proliferation and nuclear receptor signaling pathways were strikingly different at the transcriptome relative to the proteome levels (Fig. 6A). These include the canonical pathways for a) FXR/RXR and LXR/RXR Activation, which were modestly enriched bidirectionally without clear patterning in the transcriptomic data, but strongly unidirectionally decreasing in the proteomic data for M-B and M-D, and b) Cell Cycle Control of Chromosomal Replication and G2/M DNA Damage Checkpoint Regulation, for which the proteomic and transcriptomic data were similarly enriched but varied in significance according to stage. These results demonstrate multiple levels of post-transcriptional regulation and highlight the value of multi-omic analyses to establish a holistic understanding of biological changes during early hyperplasia.

Transition to dysplasia and adenocarcinoma

Large numbers of molecular (n = 5,380) and pathway (n = 631) changes were also observed during dysplasia onset (from mucosa to benign and to dysplasia, corrected p-value < 0.05 for molecules; nominal for pathways). In most cases (ie. transcriptome, metabolome and lipidome) these changes could be observed between benign and dysplastic polyps. For proteomics the limited number of benign samples (due to limited sample size) resulted in the identification of fewer novel unique proteins. Many of the pathways evident during benign polyp formation were evident in the dysplasia polyps as well (cell cycle/proliferation). However, several pathways including nuclear receptor signaling and tumor microenvironment pathways became specifically activated during dysplasia formation (Figure 6A). At the compound level, we captured significant alterations of several key lipid signaling pathways (e.g. triglyceride catabolism and metabolism and triacylglycerol degradation) (Figure 6B). In addition, we observed dysregulation of some amino acids such as cysteine metabolism (FDR= 6E-04), a larger number

of alterations of nucleotide metabolism (pyrimidine metabolism; FDR= 2.9E-07) and mitochondrial signaling (FDR=9.4E-05) as well as significantly higher regulation of inflammatory pathways (arachidonic acid; FDR=4E-04/ immune system; FDR=7.6E-07) at mucosa to dysplastic polyp transition compared to mucosa to benign polyp. Of note, EP300, a histone acetyltransferase that regulates gene expression [33], is decreased in benign polyps but increased in polyps with dysplasia suggesting a significant transcription activity shift during dysplasia formation (Fig 6A). Overall, these results indicate alterations of key molecules and pathways during dysplasia.

Although our focus was on early cancer formation, we further investigated paired adenocarcinomas from three patients. Two were preserved using formaldehyde fixation and thus were not compatible with proteomic, metabolomic or lipidomics assays. Transcriptome analysis of adenocarcinomas compared with dysplastic polyps revealed significantly enriched pathways involved in cancer signaling and formation. Similarly cellular stress and signalling pathways were highly altered, including DNA double-strand break repair, GP6 signaling pathway, ATM signaling, etc. These pathways were similarly enriched in sporadic colorectal cancers profiled within TCGA.

Early steps of cancer progression

The pairwise comparisons above revealed substantial changes between benign polyps and normal tissue, as well as alterations in dysplastic polyps and adenocarcinoma. To systematically identify molecular and pathway-level changes that occur across stages of malignancy, we examined seven distinct patterns of dynamic change across stages of malignant progression and report their representative trajectories, number of significant molecules, and key pathways (Fig 5D, Supplemental Data 7). For example, Pattern 1 shows a continuous increase/decrease across the entire malignancy spectrum. The *Wnt2* transcript shares this pattern which exhibited a progessive increase in expression in the transition from mucosa through adenocarcinoma, whereas *C2orf88* and *CFD* decreased. *Wnt2* encodes secreted signaling proteins involved in the Wnt signaling pathway, which we detected as significantly enriched in both the transcriptomic and proteomic data. C2orf88 (chromosome 2 open reading frame 88), which exhibited continuously decreasing expression, was previously reported as the one of the most significantly differentially downregulated genes in CRCs[34]. Pattern 2 includes molecules that increase during benign and dysplastic polyp formation, but subsequently plateaued, including *NKD1, ACSL* and *FOSL1*. *NKD1* is a negative regulator of the Wnt signaling pathway [35–37] that has previously been reported to be up-regulated in colorectal adenomas [38]. ACSL6 has also been found to be overexpressed in colon cancer [39] and has been investigated as a long chain acyl-CoA synthetase target for cancer therapy [40]. Pattern 3 includes pathways associated with tumorigenesis and cancer cell mobility that were altered in benign polyps and in adenocarcinomas but were not significant when comparing benign and dysplastic polyps. Finally, several molecules/pathways were associated with either an increase or a decrease in the dysplastic polyp through adenocarcinoma phase (Pattern 4, “dysplastic associated molecules"). Genes altered in pattern 4 include the taurine transporter, SLC6A6, the tumor suppressor LGALS2, BTNL8 an immune response-related gene, and metallothionein (MT) protein

family members such as MT1G. Intriguingly, MT1G was previously reported to be associated with CRC prognosis[41]. Other patterns of change (Patterns: 5, 6 and 7) reveal molecules and pathways that change at the normal mucosa to benign polyp transition, benign polyp to dysplasia and dysplasia to adenocarcinoma. Collectively, these results indicate specific pathways are variably altered across normal, dysplastic and malignant tissues.

DNA methylation changes during early tumorigenesis

Finally, we examined DNA methylation in eight samples profiled via whole genome bisulfite sequencing (WGBS). Global hypomethylation increased across all transitions, from colorectal mucosa to benign and dysplastic polyps to adenocarcinoma (Figure 5E); in particular, the fraction of partial methylation domains (PMD) increased up to 100% (Figure 5E). At the gene level, promoter hypomethylation was observed in the transition from mucosa to benign polyp for *WNT2*, *TGIF1* and *SOX9* (Figure 5F), which are known to be involved in Wnt signaling [42,43] , TGF-beta signaling [44] and CRC metastases [45] respectively. For these three genes, hypomethylation increased with each transition, as did RNA expression. In contrast, *PDGFD,* a gene involved in RAS signaling [46], exhibited increased hypermethylation and decreased expression as the tissue stage advanced toward adenocarcinoma. This is supported by the finding that PDGFRB protein levels decreased significantly (FDR < 0.01) in benign and dysplastic polyps relative to mucosa and by changes in ceramides (a key component of sphingolipid metabolism), which were significantly downregulated (FDR < 0.05) in both polyps relative to mucosa and associated with the PDGF family (Figure 7A). These results are consistent with WGBS data in advanced cancers from TCGA (Figure 5G)[47] and indicate that DNA methylation changes occur early during CRC formation, prior to observable dysplasia.

Discussion

We performed a comprehensive multi-omic (genome, transcriptome, proteome and metabolome, lipidome) analysis of FAP polyps (and adjacent tissues and adenocarcinomas), which represent a valuable model of colorectal tumorigenesis. These results reveal early clonal spreading as well as molecular and cellular pathways associated with hyperplasia. In FAP patients, even “normal” mucosa harbors somatic mutations with multiple driver alterations detected in the mucosa from two cases (Supplementary Figure 1A). In particular, patient A001 harbored mutations in *KMT2C* and patient G exhibits mutations in *HLA-B* and *APC* (Supplementary Data 2). In contrast, analysis of healthy tissue from other organs profiled within HubMAP (data not shown) revealed minimal somatic alterations, consistent with deep targeted sequencing of colon crypts of 50-60 yr old donors which revealed 1% mutations in normal crypts [48]. Collectively, these data suggest that the germline loss of function of one APC allele is permissive for subsequent alterations.

Analysis of cancer driver gene frequency and clonality in FAP polyps indicates that KRAS mutations occur early in oncogenesis as many benign polyps contain these mutations. This is often followed by *FBXW7* and secondary *APC* mutations. Analysis of 2,658 sporadic cancers yields similar results [48]

indicating both that FAP is a useful model for CRC and that these mutations occur early, often prior to dysplasia formation. Many, but not all, samples harbored a second APC mutation. It is possible that those samples lacking coding mutations harbored *APC* promoter mutations, however expression differences were not apparent between cases with or without somatic *APC* mutations (data not shown). Rather, mutations in *KRAS*, *FBXW7*, *SMAD4* and *TP53* were common, suggesting that secondary mutations in other pathways, on the backdrop of germline *APC* loss can promote tumorigenesis.

Drawing on the numerous and spatially diverse samples spanning the continuum of malignancy within multiple FAP donors, we investigated their developmental and evolutionary histories based on patterns of somatic alterations. Of note, we detected shared mutations across polyps throughout the colon, and it is upon this background that thousands of independent lesion-specific mutations subsequently arose. We found that many of these early mutations which are “shared” throughout the colon are cancer driver events, such as hotspot mutations in KRAS (Figure 3D). These oncogenic mutations are known to elevate crypt fission rates and contribute to the fixation and spreading of clonal populations within the colon [49] . The extensive subclonal mutation sharing observed between lesions within a given patient might be explained by a model of FAP development in which multiple clones arise in the early post-zygotic colon (possibly in the endoderm layer), creating a mosaic in the maturing colon. Polyps that arise within this mosaic of clones in the FAP colon are thus composed of multiple “polyp-founders”.

Multi-omic analyses revealed numerous molecular and pathway changes in benign polyps reflecting early changes in early cancer formation (Fig 5, Fig 6). Transitions to benign polyps from mucosa were replete with cell proliferation and immune system alterations in both the transcriptome and proteome. Examination of the lipidome and metabolome revealed several metabolic pathways (e.g. arachidonic acid, amino acids, carbohydrates, nucleotides, lipids, and hormones); such pathways interface with key biological systems involved in cellular homeostasis and organismal survival such as the immune system (e.g. arachidonic acid and acetylsalicylic acid/aspirin are central to inflammation regulation). Cellular systems involved in energy production and regulation also connect with important metabolic pathways for molecules serving as the preferred energy supply for cell growth and survival, including carbohydrates (e.g. Lactose Synthesis), lipids (e.g. Phosphatidylcholine Biosynthesis), and amino acids (e.g. Arginine and Proline Metabolism).

The transition of polyps to dysplasia was characterized by proteomic and transcriptomic alterations in Wnt and nuclear receptor signaling, as well as stromal and immune pathways. The observed patterns of lipid dysregulation in TAG metabolism suggests a further energy metabolism shift and putative role for TAGs as an energy source in polyp growth leading to dysplasia. Transitions to adenocarcinoma were characterized by extensive transcriptomic alterations, including upregulation of cellular stress, immune response, and cancer signaling pathways that were modestly changed in the polyp analysis. The upregulation of immune pathways is notable since immune genes were generally downregulated across polyp transitions. More generally, the immune response appears to be activated early during colon hypergrowth, likely reflecting more potent immune surveillance during early disease stages. Arachidonic acid signaling is also upregulated early during polyp formation, contributing to inflammation. While

aspirin/NSAIDs are commonly used as prophylactic treatment of FAP patients, the mechanism of action and optimal timing of this intervention is not known. Of note, we detected significantly decreased expression of the key genes in the arachidonic acid signaling pathway. One of them is PLA2 (phospholipase A2), which converts diacylglycerol or phospholipids to arachidonic acid. Another one is ALOX5 (lipoxygenase), which uses arachidonic acid to produce HPETE and also leukotriene A4 that has an established pathological role in inflammation and infection. But increased arachidonic acid signalling molecules were detected at the metabolic level, together with the dysregulation of key genes in the pathway, emphasizing the value of profiling key transitions at multiple molecular scales. In summary, this rich multi-modal dataset set advances an understanding of clonal evolution in hereditary pre-cancer and illuminates the steps involved in precancer formation and transition to malignancy at unprecedented resolution and should serve as a valuable resource for the community.

Methods

Patient Selection

Patients enrolled onto the protocol were screened and selected through the Stanford Cancer Genetics clinic, which provides genetic testing and counseling for persons at risk for hereditary cancer. Potentially eligible participants met one of the following criteria: 1). Molecular diagnosis of colorectal polyposis syndrome with a pathogenic/likely pathogenic genetic test result in the APC gene 2). Molecular diagnosis of biallelic MUTYH gene mutation 3). Clinical diagnosis of colorectal polyposis syndrome

We also screened and invited patients to participate who were seen through the Gastroenterology and Hepatology service and Adult and Pediatric Surgery who were undergoing colonoscopy, pouchoscopy or colectomy that met the above criteria.

Once a patient was identified, they were notified of their eligibility to participate in research. Over the phone, the description, risks, benefits, and alternatives of participating in the research study are described and a copy of the full consent form was sent to them via email. After giving verbal consent to participate, the clinical research coordinator met with the patient on the same day of the procedure to go over any questions they might have and to sign the consent forms.

Sample Collection

All of the tissues used in this study were procured from colons from total colectomy procedures. The patient colons were taken directly from the surgical suite to the Stanford Hospital pathology gross room where they were quickly rinsed and flayed open with a scalpel lengthwise on a room temperature cutting board typically used in a pathology gross room. Polyp-adjacent normal mucosa, polyps and adenocarcinoma tissues were carefully removed from the colon and stored in cryovial tubes and quickly placed into a filled portable liquid nitrogen tank. To record the exact location on the colon where the different tissue samples came from for patients A001, A002, A014, A015, we replaced each tissue sample with a numbered thumb-tack and took pictures of the flayed open colon. We stitched together the images

of the colon and determined the XY coordinates of each of the tissue samples in pixel units then used the diameter of the thumb-tacks as a scaling factor to convert pixel units to centimeters. Using a ceramic mortar and pestle, we pulverized the flash frozen samples in liquid nitrogen to create a fine powder or small chunks as input to each multi-omic assay. With Qiagen All-prep kits (cat. 80204) we isolated DNA and RNA. Buffy coat DNA was isolated by placing 200 ul directly into 600 µl Buffer RLT with β-ME at 1:100 dilution followed by the All-prep kit instructions.

Whole genome/whole exome genomic DNA analysis

The DNA was sequenced on the Illumina NovaSeq 6000 after library preparation using either the NEBNext DNA Library Prep Kit (cat. no. E7645S) or TruSeq Nano DNA Kit (cat. No. 20015964) following the manufacturers guidelines. The DNA extracted from FFPE was first treated with S1 Nuclease from Thermo Scientific (cat. no. EN0321) followed by NEBNext FFPE DNA Repair Mix (cat. no. M6630) then NEBNext DNA Library Prep Kit. After sequencing, using an in-house pipeline for genomic analysis [50], we mapped FASTQ files with bwa mem (v1.10,bwa mem -r 1.2 -t 40 -R) for hg38. We created a target interval list file with GATK's RealignerTargetCreator (v3.4-46, –allow_potentially_misencoded_quality_scores) and IndelRealigner (v3.4-46, -compress 5) with Mills_and_1000G_gold_standard.indels and dbSNP138 indels as references. To reduce technical errors by the sequencer, called bases were recalibrated with BaseRecalibrator (v3.4-46) against the previous indel references and the SNPs from dbSNP138. PrintReads (v3.4-46, -DIQ –emit_original_quals -BQSR *BaseRecalibrator-output*) outputs well-formed reads with the bases recalibrated. Finally, to create the final .bam file, duplicate reads are marked with MarkDuplicates (Picard v2.18.7, REMOVE_DUPLICATES=false VALIDATION_STRINGENCY=LENIENT ASSUME_SORTED=true).

Somatic SNV and Indel calling, CNV detection, and Cancer Cell Fraction estimation.

Somatic single nucleotide variants (SNVs) and small insertion deletions (Indels) were detected using Mutect2 (GATK v.4[51] in multi-sample mode per patient. "-normal" was the .bam file from the buffy coat. The –germline-resource was the af-only-gnomad.hg38.vcf. In order to filter out DNA damage artifacts from FFPE, we ran Mutect2 with the argument –f1r2-tar-gz then used LearnReadOrientationModel and MergeMutectStats. Finally, the mutations were filtered using FilterMutectCalls and the following arguments: -stats ${PT}.merged.stats , –orientation-bias-artifact-priors ${PT}-read-orientation-model.tar.gz , –max-events-in-region 30, –max-alt-allele-count 2, -XL hg38-blocklist.v2.bed , -XL ENCFF356LFX.bed. The filtered mutations were left-aligned and multi-allelic mutations were split with LeftAlignAndTrimVariants. Mutations were annotated using Annovar with the following databases for hg38: refGene, clinvar_20170905, cosmic70, dbnsfp33a, cytoBand, avsnp150. Furthermore, any mutations found in the patient blood samples were also removed from any downstream steps as well. Also, mutations were considered present if supporting variant reads (altc) >=3 and the total number of reads at that variant position was >=10 reads. Mutations with CADD>=20 and "pathogenic" CLINSIG annotations and found in genes within the list of cancer driver genes from Bailey MH, et al.[20–22] were plotted.

We used Titan to detect copy number variants (CNVs) and estimate purity and ploidy. Differentially altered somatic copy number alterations (sCNAs) in individual FAP samples were identified if the copy number of the sample relative to the median ploidy of the samples's stage group (ie. mucosa, benign, dysplastic and adenocarcinoma) within a patient was >0.1 or < −0.1. The sCNA frequency is the number of times that region of the genome is altered divided by the number of samples in that stage group. To identify recurrently overlapping regions we used the GISTIC procedure in CNVRanger::populationRanges() in reciprocal overlap mode with default 0.5 threshold and assessed recurrence of overlaps using permutation testing.

Cancer cell fraction (CCF) was calculated using the R software CNAqc with default settings[23]. To characterize mutations as "clonal" or "subclonal", we first create 2 pseudo variables:

1 - Purity-adjusted depth takes the total read count at each mutation and multiplies it by the tumor-purity estimation output by Titan.

2 - CCF-adjusted altc assumes that each mutation takes place at a 1:1 karyotype for simplicity. This was calculated by taking the (CCF*p)*(purity-adjusted depth) where p=0.5. When p=.5, CCF*p produces the expected variant allele frequency (VAF) if the mutation had occurred at a 1:1 karyotype.

Finally, to characterize a mutation as clonal, we determine if the CCF-adjusted altc was within 1 single standard deviation of the binomial distribution mean of B(n,p) where n is purity-adjusted depth and p = 0.5. If the mutations CCF-adjusted altc fell outside of the standard deviation, it was considered "subclonal". Also, any CCF values above 0.85 were considered "clonal".

Additionally, for linear regression models comparing the number of mutations detected per sample to different clinical factors, it was important to normalize the number of mutations detected. First, we calculated the mean depth of coverage for all of the detected mutations per sample. Then per patient, we calculated the 25% quartile of the mean depths of coverage. Next, we normalized the number of mutations detected in each sample by determining the likelihood of the mutation being detected at a lower depth. This was done by testing if there was a greater than 90% chance of there being >= 3 alternate reads in a binomial distribution in which B(n=25% quartile of mean depth,p=(CCF/2)). Only mutations which passed this test were included in the "normalized" counts. The linear mixed effects regressions were performed with the nlme package in R with the lme() function to predict the log-tranformed number of SNVs. The patient was considered a random variable and the model was weighted by Stage (M, B, D, AdCa) using varIdent(). The models were calculated for maximum likelihood ("ML").

Maximum Parsimony Phylogenetic Tree Building

Only SNVs which passed the above detection filters were considered in building the maximum parsimony trees in R using the software phangorn and ape [52,53]. Initially, the trees were built with pratchet() with the Fitch method, iterated 1,000,000 times, and had its edge lengths adjusted with acctran(). The consistency index value was calculated using CI = CI() and the homoplasy index was HI = 1-CI. The bootstrap values

were generated by rebuilding the trees 100 times using bootstrap.phyDat(). We determined which mutations belong on each edge of the tree by using ancestral sequence reconstruction with ancestral.pars(). Thus mutations found only on terminal branches of the tree were considered “private”, the rest of the mutations found on internal branches were considered “shared”. Phylogenetic trees were plotted using plotBreakLongEdges(). Only cancer driver gene mutations with CADD>=20 or “pathogenic” CLINSIG annotations were shown on terminal branches determined through ancestral sequence reconstruction. Minor aesthetics of the figures were adjusted manually in Adobe Illustrator.

Subclonal Mutation Sharing between FAP Lesions

To quantify the amount of subclonal sharing occurring between lesions within a patient, we used a customized Jaccard similarity index (JSI); per pair of samples within the same patient, divide the number of identical subclonal mutations by the sum of private clonal mutations and identical subclonal mutations. A value of 1 means there are no differences between the samples where a value close to 0 means there are very few identical subclonal mutations and/or very many private clonal mutations distinguishing the 2 samples. To determine the physical distance between the tissue samples on the colectomy samples, we took the x-coordinate and y-coordinate in centimeters of the tissue samples from the stitched together images and simply calculated the euclidean distance in R with dist().

Relative Mutation Timing Analysis

Mutations were ordered chronologically using 2 criteria: whether the mutation was clonal or subclonal and whether the mutation was shared or private. The rationale here is that clonal mutations are more abundant in a sample and therefore arise before subclonal mutations. Similarly, if a mutation is shared between samples, it is more ubiquitous within a set of samples in a patient and therefore arose before private mutations. Per sample, using the above categories for all cancer driver mutations with CADD>=20 or “pathogenic” CLINSIG annotations in a pairwise fashion, we counted the number of times every mutation in a specific gene came before another. Then we performed Bradley Terry modelling and used the coefficients to determine the relative chronology for each mutation. We performed leave-one-out (LOO) cross validation by analyzing all the samples of interest except for one, then repeating this process *n (n*=# of samples*)* times with a different sample removed from the analysis each time. The rank order of the genes is the median of the coefficients from the LOO analysis.

Predicting Polyp Stage Based on (sub)Clonal Mutation Distribution

To determine the extent to which clonal distribution can be associated with polyp development we used the (sub)clonal CCF distributions found in the violin plots in **Figure 3D**, Supp Figs 3C,G,J,M,P to create a 1 dimensional trajectory of the development of polyps from mucosa to benign and dysplastic polyps to adenocarcinoma. We created 100 bootstraps per sample (using and randomized 60% of the data each time) and calculated multiple metrics describing the (sub)clonal distributions including standard deviation, interquartile range, median, Kolmogrov-Smirnov (stats::ks.test()) and Z-statistics ((mean1-mean2)/sqrt(mean1^2+mean^2)). We calculated these statistics for the shared and private mutations as

wholes then broken down further into shared-clonal versus private-clonal, for example. To initially select features which could help build the 1D trajectory, we first tested each of the features in a linear regression lm(stage ~ clonal.feature) in which each stage was assigned a numeric values for simplicity: 1=mucosa, 2= benign OR dysplastic polyp, 3 = Adenocarcinoma. The results of the linear regressions are plotted in the heatmap Supplemental Figure 3Y. We manually choose 2 features with high R^2 and low Residual SD (Private_clonal.median and KS.stat.Priv.v.Shar.clon) to build a linear discriminant model to learn to distinguish Normal samples from Adenocarcinoma samples only. We created a training dataset with 60% of the bootstrapped features data, and the determined that the LDA model MASS::lda(Stage.Norm.v.AdCa ~ Private_clonal.median + KS.stat.Priv.v.Shar.clon) was 91% accurate distinguishing Normal samples from Adenocarcinoma samples. We used the coefficients of the linear discriminants from this model to calculate the LD values for all of the bootstrapped sample data. The LD values were normalized between 0 and 1 for simplicity and plotted in Supplemental Figure 3Z using geom_density() from ggplot2.

Computational modeling of mono- vs poly-clonal polyp origins

We model polyp development by simulating a stochastic birth-death process.We first simulate the growth of an embryonic “colon” from a single progenitor cell where each cell divides at probability b and dies at probability d at each cell generation (b+d=1). Here b=0.6 was used, corresponding to a mean exponential growth rate r=log(2*0.6)≈0.18. Under a neutral evolution model, all mutations are functionally neutral that occur at rate of u in the genome per cell division where u=3 was used which corresponds to 10^{-9} per base pair per cell division. In a model that incorporates selection, in addition to neutral mutations, beneficial mutations are assumed to occur at rate of u_b=10^{-4} in the genome per cell division while the selection coefficient s=0.1. Thus the birth probability for cells with beneficial mutations is b(1+s). We next simulate the growth of individual “polyps” each initiated from a single (monoclonal) or a group (n=20 or 40; polyclonal) of progenitor cells randomly sampled from the embryonic “colon” at the size of 10^6 cells by assuming relatively early initiation of polyps. The “polyps” are grown under the same conditions and evolution model as the “colon” except each polyp can only grow to be 10^5 cells. For each patient, 25 “polyps” were simulated while the median CCF (y-axis) for a number of mutations present in x polyps (x-axis) were plotted (Supplementary Figure 4N).

Transcriptomic analysis

Bulk RNA-seq data was generated for 99 human colon biopsies from 6 FAP patients. We optimized tissue preservation and processing methods to ensure high quality RNA for sequencing determined by RIN scores and developed an SOP for tissue collection and processing of human colon tissues. Flash frozen tissues were lysed in RLT buffer and processed using the Qiagen AllPrep Kit for bulk RNA extraction and sequencing. Samples were sequenced to 50 million paired-end read pairs.

During quality control assessment, data quality was assessed using a RIN score cutoff of 6.0. We first performed read mapping and alignment using STAR[54] and transcript quantification with RSEM[55]. Technical replicates were collapsed via DESeq2[56]. We performed differential gene expression analysis on sample stage using DESeq2[56] (with adaptive log2 fold change shrinkage) controlling for batch, patient, colon location (ascending, transverse, descending, and rectum), and the top two surrogate variables inferred by surrogate variable analysis (SVA)[57]. Using these expression profiles, we subsequently performed pathway analysis. For pattern analysis and visualization, batch adjusted expression profiles were produced by running LIMMA[58] in R with log2 counts per million transformed expression and the same design matrix used in the differential analysis, albeit without batch as a covariate. The transcriptomic differential expression contrast results and the R code to produce it are included in Supplemental Data 3.

Proteomic analysis

Sample Preparation

Tissue samples were cut into small pieces on ice and further disrupted using beat beating and sonication in lysis buffer (6 M guanidine, 10mM TCEP, 40mM CAA, 100mM Tris pH 8.5). The supernatant was collected and heated at 95 $^{\circ}C$ for 5min. After protein reduction and alkylation, protein concentration was measured using the BCA kit (ThermoFisher). Protein extract was digested using LysC (1:100 protease to protein ratio) for 2 hours at room temperature followed by trypsin (1:50) digestion overnight at 37□C. Peptides were cleaned up using Waters HLB column and subsequently labeled using TMT 10Plex (ThermoFisher) in 100mM TEAB buffer according to manufacturer's recommendations. An equal amount of proteins from each tissue were pooled together as a reference sample.

TMT Experimental Design

In this study, we used TMT10plex which can label up to 10 samples in one experiment. We randomized tissue samples so that each TMT10plex consists of an assortment of tissues. To facilitate cross-tissue comparison and reduce the technical variation among mass-spectrometry runs, pooled reference samples were added into each TMT10plex experiment. Equal amounts of 9 samples and common reference samples were multiplexed into one TMT10plex run. We also performed two technical replicates for all samples. In the replicate, samples have the same TMT labels as in the initial run but they were re-randomized to mix with different samples in each TMT10plex run.

Two Dimensional Liquid Chromatography Separation

We used the Waters online nano 2D LC system for fractionation using approximately 15ug of multiplexed sample. Peptides were separated by reverse-phase chromatography at high pH in the first dimension, followed by an orthogonal separation at low pH in the second dimension. In the first dimension, the mobile phases were buffer A: 20mM ammonium formate at pH10 and buffer B: Acetonitrile. Peptides were separated on an Xbridge 300µm x 5 cm C18 5.0µm column (Waters) using 12 discontinuous steps

of buffer B at 2 µl/min flow rate. In the second dimension, peptides were loaded to an in-house packed 75µm ID/15µm tip ID x 25cm Sepax GP-C18 1.8µm resin column with buffer A (0.1% formic acid in water). Peptides were separated with a linear gradient from 5% to 30% buffer B (0.1% formic acid in acetonitrile) at a flow rate of 600 nl/min in 180 min. The LC system was directly coupled in-line with an Orbitrap Fusion Lumos (Thermo Fisher Scientific).

Mass Spectrometry Data Acquisition and Analysis

The Orbitrap Fusion was operated in a data-dependent mode for both MS2 and MS3. MS1 scan was acquired in the Orbitrap mass analyzer with resolution 120,000 at m/z 400. Top speed instrument method was used for MS2 and MS3. For MS2, the isolation width was set at 0.7 Da and isolated precursors were fragmented by CID at a normalized collision energy (NCE) of 35% and analyzed in the ion trap using “turbo” scan. Following the acquisition of each MS2 spectrum, a synchronous precursor selection (SPS) MS3 scan was collected on the top 5 most intense ions in the MS2 spectrum. SPS-MS3 precursors were fragmented by higher energy collision-induced dissociation (HCD) at an NCE of 65% and analyzed using the Orbitrap at a resolution of 60,000. Each sample was run again on another Orbitrap Fusion in the same lab with the exact same settings for technique replicates.

We used SEQUEST in ProteomeDiscoverer 2.1 (ThermoFisher Scientific) for protein identification. Raw files from 12 fractions of each sample were combined together for a single search against the UniProt human proteome database. Mass tolerance of 10ppm was used for precursor ion and 0.6 Dalton for fragment ions. The search included cysteine carbamidomethylation as a fixed modification. Peptide N-terminal and lysine TMT 10plex modification, protein N-terminal acetylation and methionine oxidation were set as variable modifications. Up to two missed cleavages were allowed for trypsin digestion. The peptide false discovery rate (FDR) was set as <1% using Percolator. For protein identification, at least one unique peptide with a minimum 6 amino acid length was required. For protein quantitation, only unique peptides with precursor ion isolation purity >50% were used. Peptides passing the criteria were summed to represent protein abundance.

Untargeted Metabolomic and Targeted Lipidomic analysis

Sample Preparation. Comprehensive profiling of complex lipids and metabolites were performed on 77 colon tissue samples from 6 patients (Supplemental Data 1). Roughly 30 mg of frozen tissue were homogenized in 500 µl ice-cold methanol by bead beating (MP bioscience cat# 6913-100, Solon, OH) at 4°C (2 x 45 s). Metabolites and complex lipids were extracted using a biphasic separation with cold methyl tert-butyl ether (MTBE), methanol and water. Briefly, 1 ml of ice-cold MTBE was added to 300 µl of the homogenate spiked-in with 40 µl deuterated lipid internal standards (Sciex, cat# 5040156, lot# LPISTDKIT-102). The samples were then sonicated (3 x 30 s) and agitated at 4°C for 30 min. After addition of 250 µl of ice-cold water, the samples were vortexed for 1 min and centrifuged at 14,000 g for 5 min at 20°C. The upper organic phase contains the lipids, the lower aqueous phase contains the metabolites and the proteins are precipitated at the bottom of the tube. For quality controls, 3 reference

plasma samples (40 µl plasma), twof normal colon tissue and one colonic polyps as well as one control lacking any sample (i.e. blank) were processed in parallel.

1) Metabolites: Proteins were further precipitated by adding 700 µl of 33/33/33 acetone/acetonitrile/methanol spiked-in with 15 labeled metabolite internal standards to 300 µl of the aqueous phase and 200 µl of the lipid phase and incubating the samples overnight at -20°C. After centrifugation at 17,000 g for 10 min at 4°C, the metabolic extracts were dried down to completion and resuspended in 100 µl 50/50 methanol/water.

2) Complex lipids: 700 µl of the organic phase was dried down under a stream of nitrogen and resolubilized in 200 µl of methanol for storage at -20°C until analysis. The day of the analysis, samples were dried down, resuspended in 300 µl of 10 mM ammonium acetate in 90/10 methanol/toluene and centrifuged at 16,000 g for 5 min at 24°C.

Data acquisition. Metabolite extracts were analyzed using a broad-spectrum untargeted LC-MS platform as previously described [59] while complex lipids were quantified using a targeted MS-based approach [60].

1) Untargeted Metabolomics by Liquid Chromatography (LC)-MS. Metabolic extracts were analyzed four times using HILIC and RPLC separation in both positive and negative ionization modes. Data were acquired on a Thermo Q Exactive HF mass spectrometer for HILIC (Thermo Fisher Scientific, Bremen, Germany) and a Thermo Q Exactive mass spectrometer for RPLC (Thermo Fisher Scientific, Bremen, Germany). Both instruments were equipped with a HESI-II probe and operated in full MS scan mode. MS/MS data were acquired on quality control samples (QC) consisting of an equimolar mixture of all samples in the study. HILIC experiments were performed using a ZIC-HILIC column 2.1 x 100 mm, 3.5 µm, 200Å (Merck Millipore, Darmstadt, Germany) and mobile phase solvents consisting of 10 mM ammonium acetate in 50/50 acetonitrile/water (A) and 10 mM ammonium acetate in 95/5 acetonitrile/water (B). RPLC experiments were performed using a Zorbax SBaq column 2.1 x 50 mm, 1.7 µm, 100Å (Agilent Technologies, Palo Alto, CA) and mobile phase solvents consisting of 0.06% acetic acid in water (A) and 0.06% acetic acid in methanol (B). Data quality was ensured by (i) injecting 6 and 12 pool samples to equilibrate the LC-MS system prior to running the sequence for RPLC and HILIC, respectively, (ii) injecting a pool sample every 10 injections to control for signal deviation with time, and (iii) checking mass accuracy, retention time and peak shape of internal standards in each sample.

2) Targeted Lipidomics using the Lipidyzer Platform. Lipid extracts were analyzed using the Lipidyzer platform that comprises a 5500 QTRAP system equipped with a SelexION differential mobility spectrometry (DMS) interface (Sciex) and a high flow LC-30AD solvent delivery unit (Shimazdu, Columbia, MD). Briefly, lipid molecular species were identified and quantified using multiple reaction monitoring (MRM) and positive/negative ionization switching. Two acquisition methods were employed covering 13 lipid classes; method 1 had SelexION voltages turned on while method 2 had SelexION voltages turned off. Data quality was ensured by i) tuning the DMS compensation voltages using a set of lipid standards (cat# 5040141, Sciex) after each cleaning, more than 24 hours of idling or 3 days of consecutive use, ii)

performing a quick system suitability test (QSST) (cat# 5040407, Sciex) before each batch to ensure acceptable limit of detection for each lipid class, and iii) triplicate injection of lipids extracted from a reference plasma sample (cat# 4386703, Sciex) at the beginning of the batch.

Data Processing

Data was acquired in two separate batches and batch effect was controlled by running 3 samples in common in both batches.

Metabolomics: Data from each mode were independently analyzed using Progenesis QI software (v2.3) (Nonlinear Dynamics, Durham, NC). Metabolic features from blanks and those that didn't show sufficient linearity upon dilution in QC samples ($r<0.6$) were discarded. Only metabolic features present in >2/3 of the samples were kept for further analysis. Median normalization was applied to correct for differential starting material quantity. Missing values were imputed by drawing from a random distribution of low values in the corresponding sample. Batch effect was corrected by the ComBat model using the dbnorm package [61,62] Data quality post-normalization was verified by ensuring clustering of 3 biological replicates analyzed in two batches on a principal component analysis plot (Supplemental Figure 5A). Data from each mode were merged and 10,520 metabolic features were annotated as follows. Peak annotation was first performed by matching experimental m/z, retention time and MS/MS spectra to an in-house library of analytical-grade standards. Remaining peaks were identified by matching experimental m/z and fragmentation spectra to publicly available databases including HMDB, MoNA and MassBank using the R package 'metID' (v0.2.0). Metabolites were reported if the similarity score was above 0.65. We used the Metabolomics Standards Initiative (MSI) level of confidence to grade metabolite annotation confidence (level 1 - level 3). Level 1 represents formal identifications where the biological signal matches accurate mass, retention time and fragmentation spectra of an authentic standard run on the same platform. For level 2 identification, the biological signal matches accurate mass and fragmentation spectra available in one of the public databases listed above. Level 3 represents putative identifications that are the most likely name based on previous knowledge. In total, 1,157 metabolites were identified and their abundances were reported as spectral counts.

Targeted Lipidomics: Lipidyzer data were reported by the Lipidomics Workflow Manager (LWM, v1.0.5.0) software which calculates concentrations for each detected lipid as average intensity of the analyte MRM/average intensity of the most structurally similar internal standard (IS) MRM multiplied by its concentration. Lipids detected in less than 2/3 of the samples were discarded and missing values were imputed by drawing from a random distribution of low values class-wise in the corresponding sample. Median normalization (excluding TAG and DAG) was applied to correct for differential starting material quantity. Batch normalization was performed using quality control reference plasma samples run in both batches. Data quality post-normalization was verified by ensuring clustering of 3 biological replicates analyzed in two batches on a principal component analysis plot (Supplemental Figure 5B). We detected 514 individual lipid species belonging to 12 classes (e.g. CE, CER, DAG, FFA, HCER, LCER, LPE, LPC, PC, PE, SM, TAG) and their abundance was reported as concentrations in nmol/g.

Differential Analysis. Alterations in FAP metabolome and lipidome were separately determined applying a generalized linear mixed model (i.e.Glmer) from lmerTest package (version 3.1.3) in R environment (version 4.0.2). Differential measurements were performed on 77 samples from 6 patients. Both fixed effects (i.e. gender and age) and random effects (i.e. patient ID) are specified via the model formulated with verbose=2, nAGQ=9 with bobyqa optimizer for controlling parameters with a set maximum of 100,000 iterations. Binary logistic regression was considered to predict response distribution (i.e. FAP phenotypes; benign and dysplastic polyps relative to mucosa). Both single (include all variables) and anova models (include and exclude metabolite/lipid) were applied. Considering stringent criteria, the highest detected *p*-value per-hit is considered as a level of significance. Data was log2 transformed prior analysis and thus estimated coefficient per-variable is fairly the log2 fold change weighted for mean variance. Thereby, for being consistent with other omics data, estimated coefficient of metabolic feature (metabolites/lipids) reported as log2 fold change.

Pattern analysis for significant molecules

In order to investigate the trajectory of each transition in cancer progression, we classify the significantly differential molecules to seven different patterns based on whether there are significant changes within each transition (e.g. from mucosa to benign, from benign to dysplasia, and from dysplasia to adenocarcinomas). Each of these seven patterns are further divided into the increasing and decreasing trends.The numbers of the significantly differential molecules were counted and investigated. Pathway enrichment analysis by using IPA was conducted for each pattern, and the key pathways are highlighted for each pattern.

Pathway enrichment analysis

For the proteomic and transcriptomic data, we used the Ingenuity pathway analysis (IPA, QIAGEN) (https://www.qiagenbioinformatics.com/products/ingenuitypathway-analysis) platform to conduct pathway enrichment analysis by using differentially expressed genes and circulating proteins with FDR < 0.01 and fold change (not shrunk) at least 50% increasing or decreasing. All the detected transcripts and proteins were used as background. Significance levels of pathways were determined by the hypergeometric test (one-sided) in IPA. To determine functional mechanisms behind the significant changes at the level of metabolome and lipidome, we performed metabolite set enrichment analysis (MSEA) and over-representation analysis, respectively. Functional annotations were performed against databases such as KEGG, SMPDB, BioCarta and HumanCys. The obtained p- values from enrichment analysis were corrected for multiple hypotheses using the Benjamini-Hochberg method and pathways with FDR below 0.05 were considered significant.

For lipidomics and metabolomics, comprehensive pathway analysis was performed for early transitions (M-B, M-D, B-D). Enrichment analysis and pathway curation obtained using significant hits by mapping of either data matrix (metabolite set) or KEGG identifier (lipid set) against several pathway datasets such as KEGG, BioCarta, SMPDB, Reactome, Wikipath and HUMCYC. In addition, incorporating interaction between different biological layers, topology-based pathway analysis performed using metabolites/lipids

KEGG ID [29]. Overall, in metabolism centric approaches, we captured over 360 pathways significantly altered in the metabolome/lipidome of polyps during transitions. Only a subset of pathways that are mainly represented by MSEA [27] and ORA [29] are plotted. Significant criteria of FDR< 5% considered to report enriched pathways. Biological functions determined by the identical compounds (lipids/metabolites) were merged and reported by the highest enriched significant level.

Multi-omic integration analysis at pathway level

The significantly enriched pathways from different -omes were curated and investigated systematically by checking across contrasts and across omes. Based on the functionality and diversity of enriched pathways, we classified the key significant pathways to the following categories: (1) cell cycle, proliferation, and development; (2) cellular stress and injury; (3) organismal growth and development; (4) cancer signaling; (5) immune response; (6) nuclear receptor signaling; (7) disease-specific pathways; (8) tumor microenvironment for transcriptomics and proteomics enrichment results. We also divide the enriched pathways in metabolomics and lipidomics data to (1) amino acids; (2) carbohydrates; (3) lipids; (4) energy-related; (5) lipid mediators (eicosanoids synthesis); (6) neurotransmitter and hormone metabolism; (7) nucleic acids; (8) others. Pathway directions were calculated by using the median of fold change values for the significant molecules in the pathway.

Multi-omic network construction

To connect the detected significant molecules and pathways, we selected six key pathways as the seed pathways and overlaid with the colorectal cancer-relevant molecules in the IPA database and constructed the correlation network. The connections were built based on the prior knowledge and the key molecules/pathways were highlighted. The directions of molecules and pathways were calculated as follows:

IPA Canonical pathways enriched in multi-omic data seeded with an IPA gene set ‘Organismal Abnormalities in Colon’ and colored by the mean log2 fold change between proteomic and RNA-Seq contrasts that agree in directionality and are significant at FDR 1%. Shown below is the transition of Mucosa to Benign.

Landscape of immune response pathways

The significantly enriched pathways that are related to immune response were extracted and examined. The presentative top significant immune response pathways were highlighted within each immune cell type. The connections were built based on the Ingenuity Pathway Analysis (IPA) database and literature mining. The key lipid mediators in immunity were highlighted in the boxes. The illustration of the landscape of immune response pathways was conducted in BioRender platform (BioRender.com).

Declarations

This project which meets the definition of Human Subjects Research was conducted at Stanford University with the approval of the Stanford Institutional Review Board (protocol 47044). This institution is in compliance with requirements for protection of human subjects, including 45 CFR 46, 21 CFR 50 and 56, and 38 CFR 16.

Acknowledgments

We would like to acknowledge Ido Laish, MD (Gastroenterology Institute, Sheba Medical Center, Tel Hashomer, Ramat Gan, Israel) and Eden Feldman, MS for tissue curation and cataloging. This work was supported by NCI grant U2CCA233311 to MPS and JMF. S.A.N. was supported in part by the Stanford Graduate Fellowship in Science and Engineering and the NIHGRI Stanford Genome Training Program (5 T32 000044). This work is also supported by the NIH grant NIH S10OD025212. A.H. was funded in part by the Cancer Systems Biology Scholars Postdoctoral Fellowship Program (5R25CA180993-04).

Author Contributions

E.D.E. and A.H. designed the experiments, interpreted results, collected samples and helped write the paper.

A.H. isolated RNA, DNA, performed somatic mutation detection, phylogenetic analysis and Jaccard statistical analysis and served as the project manager.

S.W. performed much of the multi-omic integration and analyses.

C.H. performed RNA normalization, conducted many of the differential and pathways analyses, and worked on multi-omic integration.

N.B. analyzed the metabolomics and lipidomics data and performed many of the multi-omic analyses.

S.A.N. collected samples, generated bulk-RNA datasets and performed normalization and transcriptomic analyses.

T.K.G. performed cryosectioning, histological staining and prepared slides for obtaining pathology report on the procured tissue samples.

M.E. and B.M. Metabolomics and lipidomics sample preparation and data generation.

K.C. Supervised metabolomics and lipidomics data generation and performed metabolomics and lipidomics quality control, data processing and annotation.

W.R.B. attended surgical procedures and collected samples.

Z.H. provided guidance on subclonal sharing analyses and developed the polyp simulation script.

B.B. computationally stitched together colon images from colectomies and calculated XY coordinates for each tissue lesion.

M.E.M. and R.L. identified, consented and collected polyp samples from FAP patients.

A.K. set up the DNA analysis and pipeline and performed MultiQC.

S.L. provided advice on genomic and phylogenetic analyses.

U.L. identified FAP patients and performed colonoscopic polypectomy.

J.C., D.G.E., J.S., and T.L. performed pathologic analyses of polyp samples.

A.K. provided advice on RNA-seq data normalization and differential analysis.

C.C. oversaw the genomic analyses.

MPS. and J.F. supervised the entire project and many of the analyses.

Declaration of Interests

M.P.S. is cofounder and scientific advisor of Personalis, Qbio, SensOmics, January AI, Mirvie, Protos, NiMo, Onza and is on the advisory board of Genapsys. A.K. has affiliations with Biogen (consultant), SerImmune (SAB), RavelBio (scientific co-founder and SAB) and PatchBio (SAB).

References

1. Siegel, R. L., Miller, K. D. & Jemal, A. Cancer statistics, 2018. *CA Cancer J. Clin.* **68**, 7–30 (2018).
2. Green, E. D., Guyer, M. S. & National Human Genome Research Institute. Charting a course for genomic medicine from base pairs to bedside. *Nature* **470**, 204–213 (2011).
3. Cancer Genome Atlas Network. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330–337 (2012).
4. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature* **578**, 82–93 (2020).
5. Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *N. Engl. J. Med.* **319**, 525–532 (1988).
6. Fearon, E. R. & Vogelstein, B. A genetic model for colorectal tumorigenesis. *Cell* **61**, 759–767 (1990).
7. Phelps, R. A. *et al.* A two-step model for colon adenoma initiation and progression caused by APC loss. *Cell* **137**, 623–634 (2009).
8. Jasperson, K. W., Tuohy, T. M., Neklason, D. W. & Burt, R. W. Hereditary and Familial Colon Cancer. *Gastroenterology* **138**, 2044–2058 (2010).
9. Gerstung, M. *et al.* The evolutionary history of 2,658 cancers. *Nature* **578**, 122–128 (2020).

10. Esplin, E. D. & Snyder, M. P. Genomic era diagnosis and management of hereditary and sporadic colon cancer. *World J. Clin. Oncol.* **5**, 1036–1047 (2014).
11. Esplin, E. D., Oei, L. & Snyder, M. P. Personalized sequencing and the future of medicine: discovery, diagnosis and defeat of disease. *Pharmacogenomics* **15**, 1771–1790 (2014).
12. Rozenblatt-Rosen, O. *et al.* The Human Tumor Atlas Network: Charting Tumor Transitions across Space and Time at Single-Cell Resolution. *Cell* **181**, 236–249 (2020).
13. Becker, W. R. *et al.* Single-cell analyses reveal a continuum of cell state and composition changes in the malignant transformation of polyps to colorectal cancer. *bioRxiv* 2021.03.24.436532 (2021) doi:10.1101/2021.03.24.436532.
14. Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome Res.* **24**, 1881–1893 (2014).
15. Jasperson, K. W., Patel, S. G. & Ahnen, D. J. APC-Associated Polyposis Conditions. in *GeneReviews®* (eds. Adam, M. P. et al.) (University of Washington, Seattle, 1998).
16. Li, J. *et al.* Genomic and transcriptomic profiling of carcinogenesis in patients with familial adenomatous polyposis. *Gut* **69**, 1283–1293 (2020).
17. Martincorena, I. *et al.* Universal Patterns of Selection in Cancer and Somatic Tissues. *Cell* **171**, 1029–1041.e21 (2017).
18. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
19. Ried, T. *et al.* The landscape of genomic copy number alterations in colorectal cancer and their consequences on gene expression levels and disease outcome. *Mol. Aspects Med.* **69**, 48–61 (2019).
20. Bailey, M. H. *et al.* Comprehensive Characterization of Cancer Driver Genes and Mutations. *Cell* **173**, 371–385.e18 (2018).
21. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2019).
22. Landrum, M. J. *et al.* ClinVar: improvements to accessing data. *Nucleic Acids Res.* **48**, D835–D844 (2020).
23. Househam, J., Cross, W. C. H. & Caravagna, G. A fully automated approach for quality control of cancer mutations in the era of high-resolution whole genome sequencing. *bioRxiv* 2021.02.13.429885 (2021) doi:10.1101/2021.02.13.429885.
24. Cross, W. *et al.* The evolutionary landscape of colorectal tumorigenesis. *Nat Ecol Evol* **2**, 1661–1672 (2018).
25. Hu, Z., Li, Z., Ma, Z. & Curtis, C. Multi-cancer analysis of clonality and the timing of systemic spread in paired primary tumors and metastases. *Nat. Genet.* **52**, 701–708 (2020).
26. Thirlwell, C. *et al.* Clonality assessment and clonal ordering of individual neoplastic crypts shows polyclonality of colorectal adenomas. *Gastroenterology* **138**, 1441–54, 1454.e1–7 (2010).
27. Xia, J. & Wishart, D. S. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res.* **38**, W71–7 (2010).

28. Kamburov, A., Stelzl, U., Lehrach, H. & Herwig, R. The ConsensusPathDB interaction database: 2013 update. *Nucleic Acids Res.* **41**, D793–800 (2013).
29. Picart-Armada, S., Fernández-Albert, F., Vinaixa, M., Yanes, O. & Perera-Lluna, A. FELLA: an R package to enrich metabolomics data. *BMC Bioinformatics* **19**, 538 (2018).
30. Samadder, N. J. *et al.* Effect of Sulindac and Erlotinib vs Placebo on Duodenal Neoplasia in Familial Adenomatous Polyposis: A Randomized Clinical Trial. *JAMA* **315**, 1266–1275 (2016).
31. Zhang, X. *et al.* Lipid levels in serum and cancerous tissues of colorectal cancer patients. *World J. Gastroenterol.* **20**, 8646–8652 (2014).
32. Mika, A. *et al.* Decreased Triacylglycerol Content and Elevated Contents of Cell Membrane Lipids in Colorectal Cancer Tissue: A Lipidomic Study. *J. Clin. Med. Res.* **9**, (2020).
33. Attar, N. & Kurdistani, S. K. Exploitation of EP300 and CREBBP Lysine Acetyltransferases by Cancer. *Cold Spring Harb. Perspect. Med.* **7**, (2017).
34. Kok-Sin, T. *et al.* Identification of diagnostic markers in colorectal cancer via integrative epigenomics and genomics data. *Oncol. Rep.* **34**, 22–32 (2015).
35. Wharton, K. A., Jr, Zimmermann, G., Rousset, R. & Scott, M. P. Vertebrate proteins related to Drosophila Naked Cuticle bind Dishevelled and antagonize Wnt signaling. *Dev. Biol.* **234**, 93–106 (2001).
36. Yan, D. *et al.* Cell autonomous regulation of multiple Dishevelled-dependent pathways by mammalian Nkd. *Proc. Natl. Acad. Sci. U. S. A.* **98**, 3802–3807 (2001).
37. Rousset, R. *et al.* Naked cuticle targets dishevelled to antagonize Wnt signal transduction. *Genes Dev.* **15**, 658–671 (2001).
38. Guo, J. *et al.* Mutations in the human naked cuticle homolog NKD1 found in colorectal cancer alter Wnt/Dvl/beta-catenin signaling. *PLoS One* **4**, e7982 (2009).
39. Rossi Sebastiano, M. & Konstantinidou, G. Targeting Long Chain Acyl-CoA Synthetases for Cancer Therapy. *Int. J. Mol. Sci.* **20**, (2019).
40. Sánchez-Martínez, R. *et al.* A link between lipid metabolism and epithelial-mesenchymal transition provides a target for colon cancer therapy. *Oncotarget* **6**, 38719–38736 (2015).
41. Hung, K.-C. *et al.* The Expression Profile and Prognostic Significance of Metallothionein Genes in Colorectal Cancer. *Int. J. Mol. Sci.* **20**, (2019).
42. Polakis, P. Wnt signaling in cancer. *Cold Spring Harb. Perspect. Biol.* **4**, (2012).
43. Huang, C.-Z. *et al.* Sox9 transcriptionally regulates Wnt signaling in intestinal epithelial stem cells in hypomethylated crypts in the diabetic state. *Stem Cell Res. Ther.* **8**, 60 (2017).
44. Guca, E. *et al.* TGIF1 homeodomain interacts with Smad MH1 domain and represses TGF-β signaling. *Nucleic Acids Res.* **46**, 9220–9235 (2018).
45. Javier, B. M. *et al.* Recurrent, truncating SOX9 mutations are associated with SOX9 overexpression, KRAS mutation, and TP53 wild type status in colorectal carcinoma. *Oncotarget* **7**, 50875–50882 (2016).

46. Zhou, S., Gao, X., Sun, J., Lin, Z. & Huang, Y. DNA Methylation of the PDGFD Gene Promoter Increases the Risk for Intracranial Aneurysms and Brain Arteriovenous Malformations. *DNA Cell Biol.* **36**, 436–442 (2017).
47. Cao, W. *et al.* Multi-faceted epigenetic dysregulation of gene expression promotes esophageal squamous cell carcinoma. *Nat. Commun.* **11**, 3675 (2020).
48. Lee-Six, H. *et al.* The landscape of somatic mutation in normal colorectal epithelial cells. *Nature* **574**, 532–537 (2019).
49. Nicholson, A. M. *et al.* Fixation and Spread of Somatic Mutations in Adult Human Colonic Epithelium. *Cell Stem Cell* **22**, 909–918.e8 (2018).
50. Sun, R. *et al.* Between-region genetic divergence reflects the mode and tempo of tumor evolution. *Nat. Genet.* **49**, 1015–1024 (2017).
51. Benjamin, D. *et al.* Calling Somatic SNVs and Indels with Mutect2. *bioRxiv* 861054 (2019) doi:10.1101/861054.
52. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
53. Paradis, E. & Schliep, K. ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
54. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
55. Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).
56. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
57. Leek, J. T. & Storey, J. D. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genet.* **3**, 1724–1735 (2007).
58. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
59. Contrepois, K., Jiang, L. & Snyder, M. Optimized Analytical Procedures for the Untargeted Metabolomic Profiling of Human Urine and Plasma by Combining Hydrophilic Interaction (HILIC) and Reverse-Phase Liquid Chromatography (RPLC)-Mass Spectrometry. *Mol. Cell. Proteomics* **14**, 1684–1695 (2015).
60. Contrepois, K. *et al.* Cross-Platform Comparison of Untargeted and Targeted Lipidomics Approaches on Aging Mouse Plasma. *Sci. Rep.* **8**, 17747 (2018).
61. Johnson, W. E., Li, C. & Rabinovic, A. Adjusting batch effects in microarray expression data using empirical Bayes methods. *Biostatistics* **8**, 118–127 (2007).
62. Bararpour, N. *et al.* DBnorm as an R package for the comparison and selection of appropriate statistical methods for batch effect correction in metabolomic studies. *Sci. Rep.* **11**, 5657 (2021).

Figures

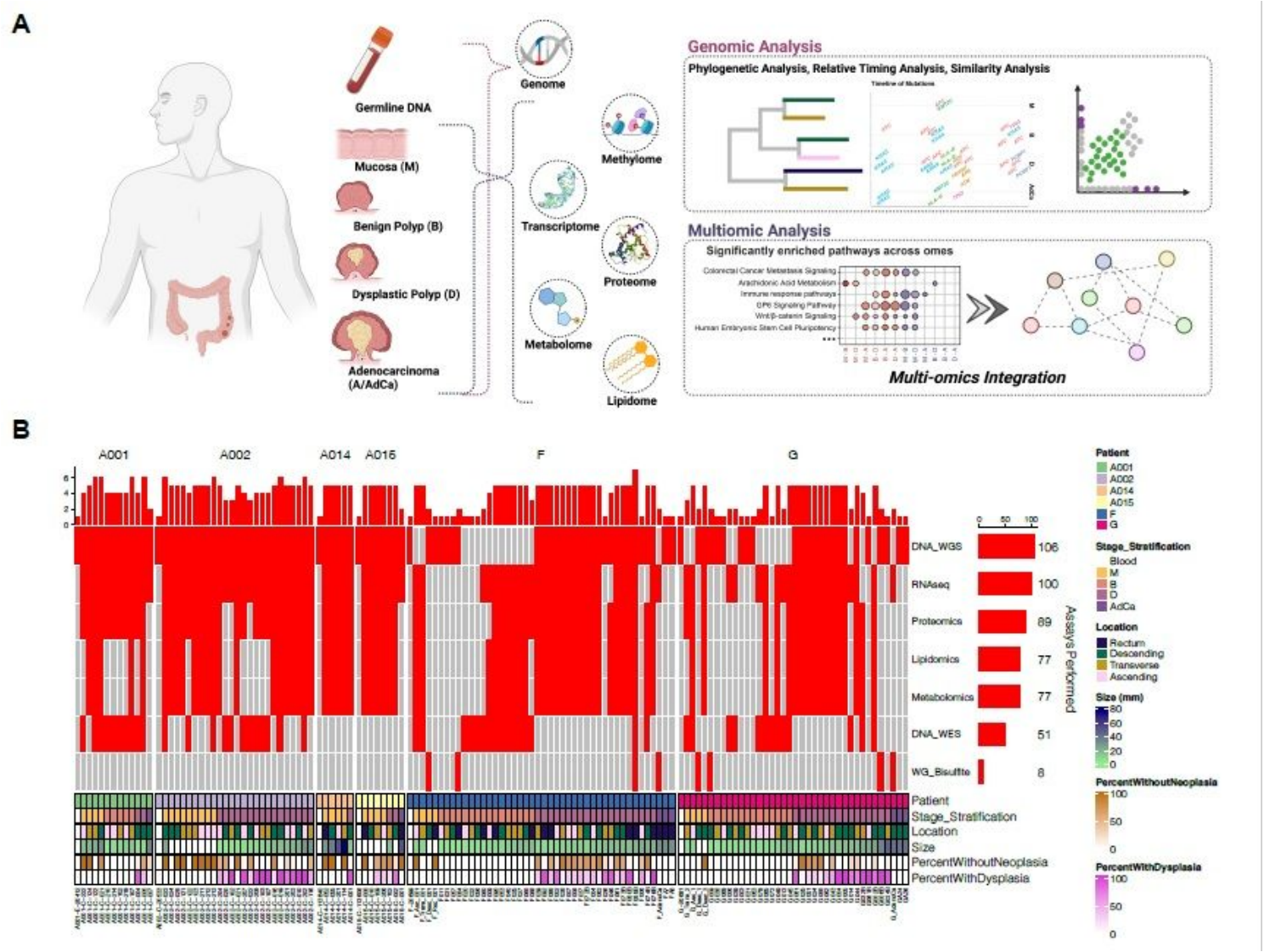

Figure 1

Overview of the multi-omic FAP study (A) Schematic of tissue and blood collections, assays and downstream analyses (genomic and multi-omic) performed in this study. (B) Summary of the clinical features of each sample and the assays performed.

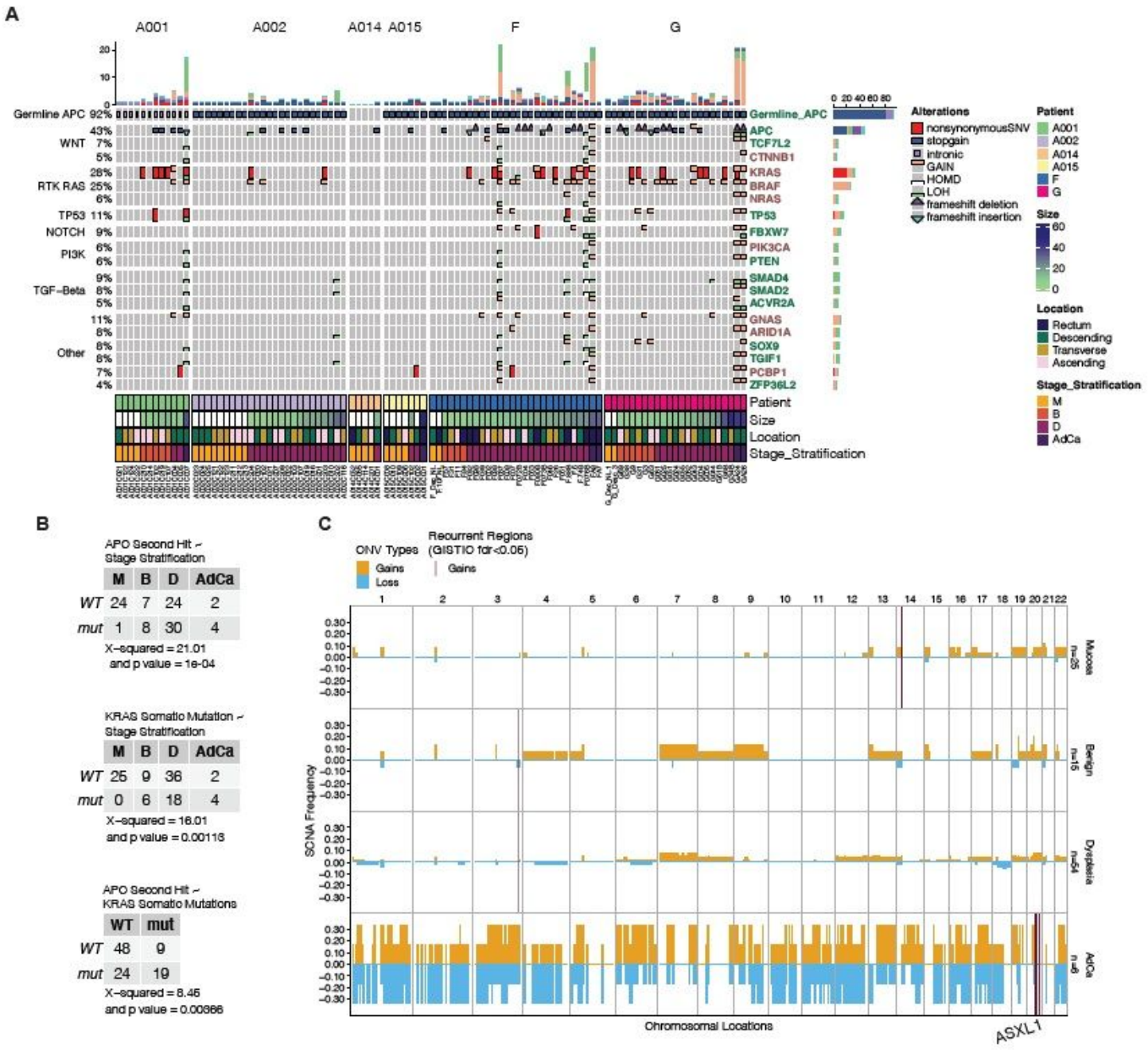

Figure 2

Mutation landscape of FAP tissues reveals Early Stage Oncogenic Events (A) Heatmap summarizing the single nucleotide variants (SNVs), small insertion/deletions (In/dels) and copy number variations (CNVs) within known colon cancer driver mutations detected using whole genome sequencing (Figure 1C) 20. The plotted SNVs and In/dels were initially filtered for functional impact (CADD_phred score >=20) or known pathogenicity (CLINSIG). Titan allele-specific copy number amplifications and loss of heterozygosity estimates were collapsed as follows: GAIN refers to allele-specific, unbalanced, and balanced copy number gains, LOH refers to hemizygous deletion, neutral and amplified loss of heterozygosity and HOMD refers to homozygous deletion. (B) Count matrices showing the numbers of

samples with wild-type and somatic mutations (SNV/In/Del/CNV) of APC and KRAS and their distribution among the different tissue stages: histologically normal mucosa, benign polyp, dysplastic polyp and AdCa. Chi-squared test values are shown below each table. (C) Somatic copy number alterations plotted along the entire autosomal genome by 1 Mbp windows. Gain and loss identification is described in the methods. Significantly recurrently overlapping regions are determined using a GISTIC procedure (see methods) and are identified in dark red vertical lines.

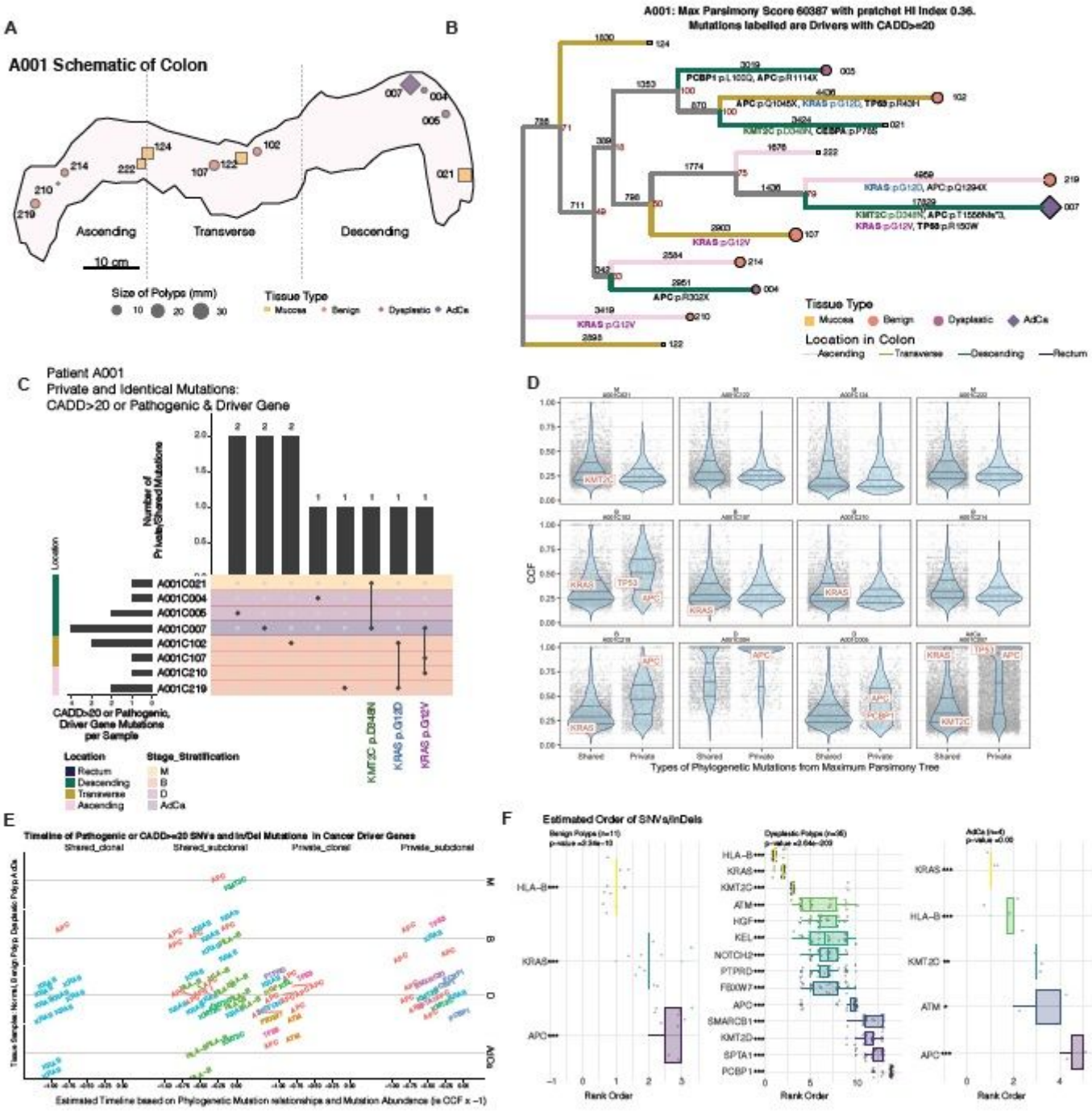

Figure 3

Identical mutations across distant FAP lesions are often subclonal and tend to occur relatively early. (A) Schematic of patient A001's colon flayed open lengthwise at time of colectomy. Each point represents a tissue sample which was collected and the whole genome sequenced. (B) Maximum parsimony phylogenetic tree constructed with SNVs and Indels. Terminal branches are labelled with mutations: occurring within cancer driver genes and are deleterious (CADD>=20) or "pathogenic" in the Clinvar database. Terminal branches are colored based on the location within the colon that the sample was procured. Terminal branch points shapes and colors represent the stage of the tissue. Black numbers above branches represent the number of mutations at that branch and red numbers at nodes are bootstrap values (the number of times out of 100 that the same clade/branch was observed when the phylogenetic tree was reconstructed on a re-sampled set of the data). (C) Upset plot showing the number of private and shared labelled mutations from 3B across all of the A001 samples. The shared mutations are shown below the plot. The location bar on the left shows the locations from the colon that the samples were taken from. (D) Violin plots showing the cancer cell fractions (CCFs) of the mutations within each sample grouped by Shared and Private mutations. The phylogenetic tree's labelled mutations (3B) are shown. (E) Relative timing of deleterious (CADD>=20) or "pathogenic" SNVS and In/Dels within cancer driver genes based on (sub)clonal and shared/private status for all samples in the cohort. (F) Rank ordering of mutations were determined using Bradley Terry modelling. A gene-by-gene matrix is created for each subset of samples and based on the 4 relative timing positions shown in 3E; per sample, we conduct pairwise "competitions" of gene A versus gene B and input the values into the aforementioned gene-by-gene matrix. We perform Bradley Terry modelling with a leave one sample out bootstrapping to produce the boxplot distributions. The p-values at the top of the plots are the likelihoods that random chance accounts for the relationship observed in the linear regression model in which the dependent variable is the rank order of each mutation and the independent variable is a one-hot encoding of the name of each gene in which the mutation occurs. The asterisks beside each gene name represent the p-values values of each gene's coefficient in the aforementioned linear model.

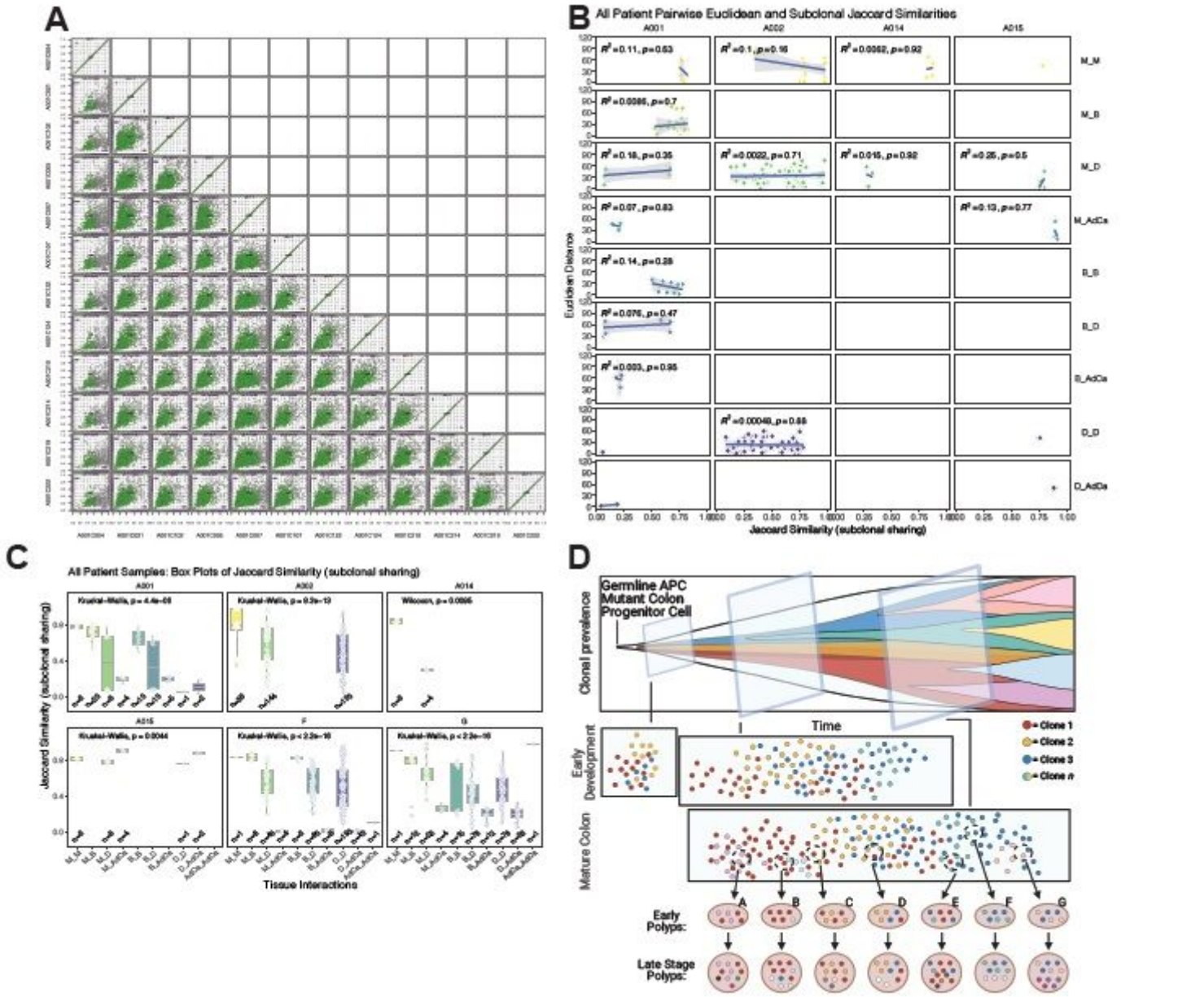

Figure 4

Abundant subclonal mutation sharing across the colon, independent of distance, suggests FAP polyps are polyclonal. (A) Using A001 as an example, this is a pairwise scatterplot where each point represents a mutation and its location is determined by the CCF. Green dots are subclonal mutations in each of the samples being compared (“shared”). Purple dots represent clonal mutations only found in one of the samples being compared (“private”). A Jaccard similarity index (JSI) is calculated by dividing the number of shared subclonal mutations by the sum of the shared subclonal and private clonal mutations between 2 samples. (B) Scatterplot comparing physical distance between tissue lesion in centimeters and JSI facetted by tissue type contrasts (binned which stages of tissues are being compared) and individual patients. Blue lines are linear models and grey bars are 95% CI. (C) Boxplots of JSI values for each type of

tissue interaction per patient. Mean comparisons was performed using Kruskal-Wallis or Wilcoxon testing. (D) Proposed schematic of how FAP patient colons develop widespread shared subclonal mutations. In the top panel, an early clonal expansion model with multiple clones is followed by additional smaller clones arising as well. The light turquoise rectangles in the center represent 1) 2D cross sections of the top clonal expansion plot and 2) simplified versions of the developing colon with the clones intermixing as they expand. The dashed-line circles indicate regions of the colon which develop into polyps and acquire new independent mutations over time.

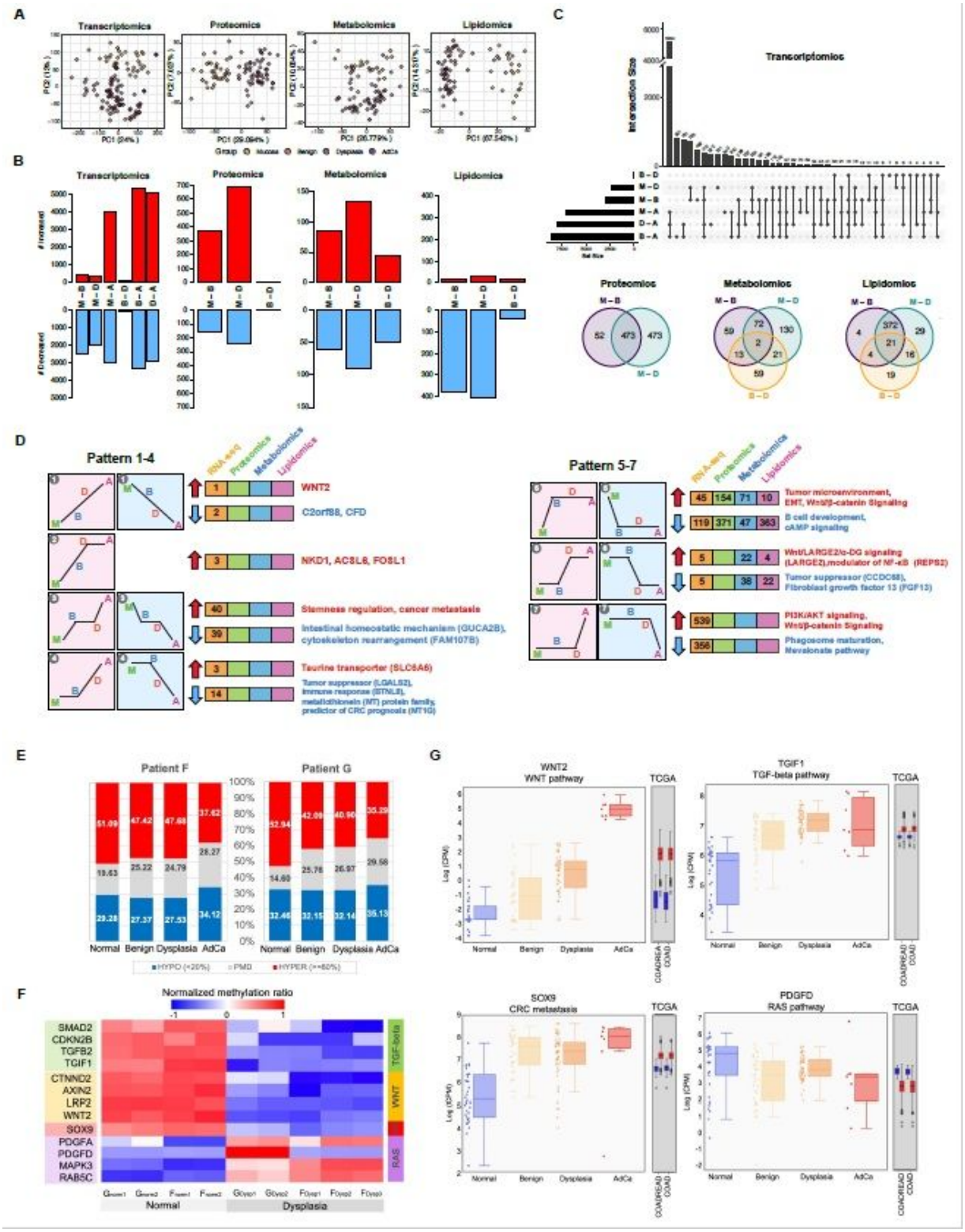

Figure 5

Multi-omic changes in FAP. (A) Principal component analysis across omic assays. Each dot represents a sample colored by different stages. (B) Bar-plot to show the numbers of significantly different molecules in each comparison in each omics dataset. (C) Comparison of significantly different molecules in each comparison. Upper panel: UpSetR plot to compare the significant gene lists in six different comparisons (from top to bottom: B - D, benign vs dysplasia; M - D, normal mucosa vs dysplasia; M - B, normal mucosa vs benign; M - A, normal mucosa vs adenocarcinoma; D - A, dysplasia vs adenocarcinoma; B - A, benign vs adenocarcinoma) in the transcriptomics dataset. Bottom panel: venn diagrams for proteomics, metabolomics, and metabolomics. (D) Pattern analysis. Schematic line charts to present the trajectory of each pattern with the increasing and decreasing trends. The values in the boxes indicate the numbers of significant molecules belonging to each pattern in each omic assay. The key molecules and selected pathways are highlighted next to each pattern. (E) Whole-Genome Bisulfite Sequencing (WGBS) applied to eight samples (normal, benign, dysplasia, and adenocarcinoma) from two patients (F and G) revealed that partial methylation increased over polyp progression. (F) Promoter methylation of major CRC-relevant genes and important pathways were manually and summarized in the heatmap. Methylation alteration between normal and dysplastic tissues suggests such epigenetic alteration is underway even before the early-onset of CRC. (G) WNT2 promoter is hypomethylated, and gene expression increases over polyp progress to adenocarcinoma. A similar pattern was found for TGIF1, which plays a key role in the TGF-beta pathway. So does SOX9, implicated in CRC metastasis. PDGFD promoter is hypermethylated, thus its abundance has decreased over the CRC tumorigenesis. WNT2, TGIF1, SOX9, and PDGFD abundance remain consistent with TCGA data.

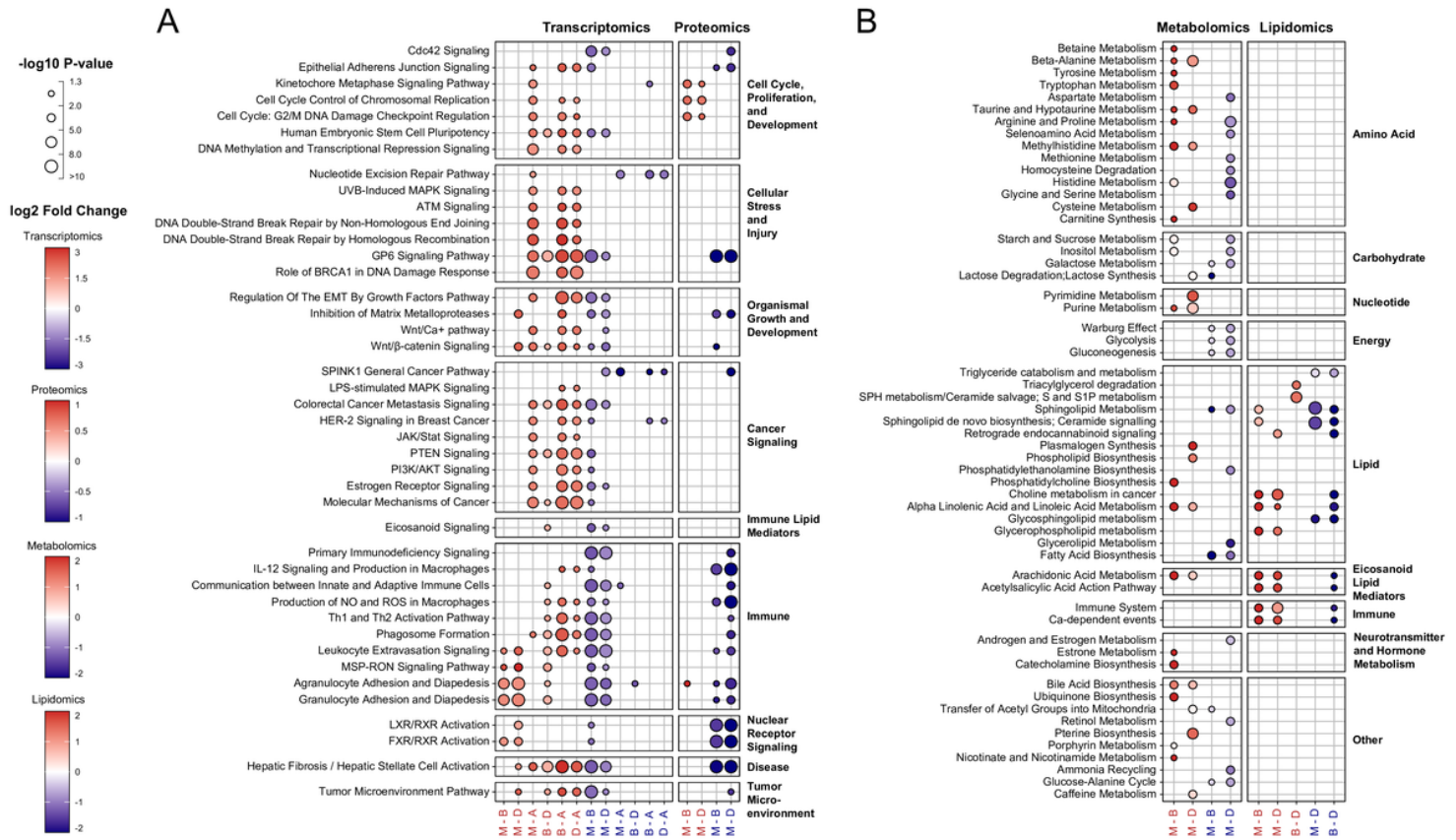

Figure 6

Multi-omic pathway changes in FAP. (A) Selected transcriptomic/proteomic pathway enrichments are shown separately by direction for each of the six contrasts based on raw IPA p-value significance at a 1% threshold to show agreement between the two omic analyses and differences. Each dot represents a (contrast, pathway, direction) triplet with size corresponding to -log10 p-value (quantized into 4 discrete sizes) and color representing the median log2 fold change (subjected to adaptive shrinkage via DESeq2 for transcripts) for molecules enriched within the pathway. Long pathways names have been abbreviated as follows: NO = Nitric Oxide, ROS = Reactive Oxygen Species, EMT = Epithelial Mesenchymal Transition, SPH = Sphingomyelin, S1P = Sphingosine-1-phosphate, and S = Sphingosine. (B) Pathway/chemical class enrichment analysis of FAP metabolome and complex lipids. Pathway direction is log2 fold change of significantly regulated molecules (p-value<5%) in the benign polyps (blue, downregulated; red, upregulated), directionality estimated relative to normal mucosa. The dot size represents pathway significance (FDR<5%). Biological functions determined by the identical compounds (lipids/metabolites) were merged.

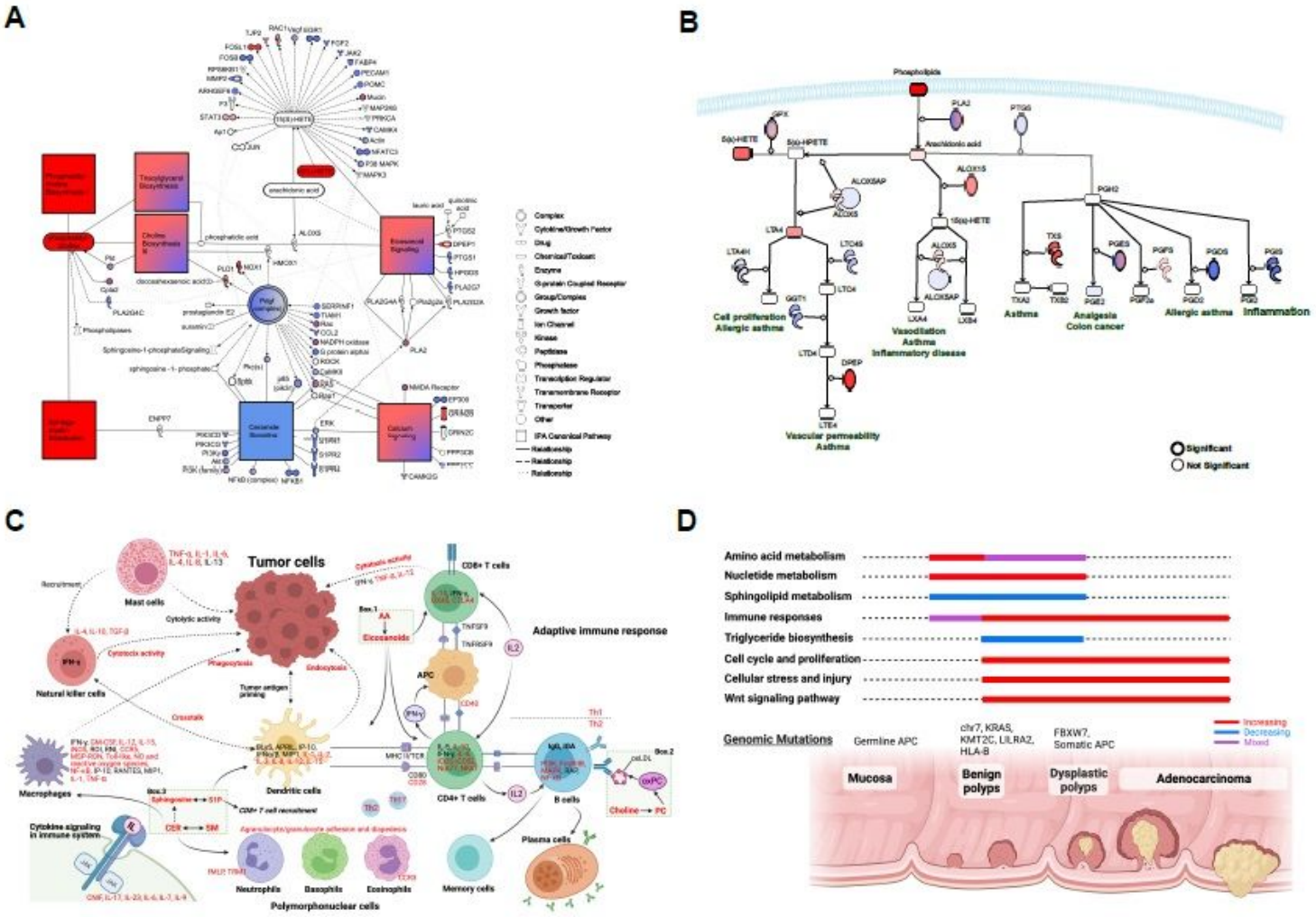

Figure 7

Summary of key biological discoveries in this study. (A) Connections between 7 enriched pathways (in at least one of the -omes), colored by log2 fold change of connected molecules from normal Mucosa to Benign. Molecules are colored according to the average of transcriptomic and proteomic log2 fold change, if they agree on direction and are both significant at FDR 0.01. Molecules significant at this contrast for both transcriptomic and proteomic data but disagree on direction from proteomic and transcriptomic contrasts are left uncolored. Molecules significant in just one -ome are colored by that -omes log2 fold change. Lipidomics and Metabolomics have 2 molecules in the network (5-Hete and phosphatidylcholine) which are both colored red for increasing and are significant at FDR 0.25. (B) Changes captured in the benign polyps relative to mucosa at the multiple molecular levels (proteins/metabolites/lipids) mapped to the eicosanoid signaling pathway: Directionality of each molecular level revealed by color code (red= up, blue= down). Color ranges were set at fold change > 1.5 or < 0.6. Significant hits are bolded. LTA4:Leukotriene A4, LTB4:Leukotriene B4, LTC4:Leukotriene C4, LTD4:Leukotriene D4, LTE4:Leukotriene E4, LXA4:Lipoxin A4, LXB4:Lipoxin B4, TXA2:Thromboxane A2, TXB2:Thromboxane B2, PGE2: Prostaglandin E2, PGF2a: Prostaglandin F2 alpha, PGD2: Prostaglandin D2, PGI2: Prostaglandin I2. (C) A panorama of immune response pathways in FAP. The representative immune signaling pathways are marked next to or in the symbols of each immune cell type. The immune response pathways that are significantly enriched in our omics data are highlighted in red. The three boxes contain the key lipid mediators involved in immunity detected in the lipidomics dataset and their chemical relationships. The connections in the map are based on the Ingenuity Pathway Analysis (IPA) database and literature. AA: arachidonic acid; PC: phosphatidylcholine; oxPC: oxidized phosphatidylcholine; oxLDL: oxidized low-density lipoprotein; S1P: sphingosine 1 phosphate; CER: ceramide; SM: sphingomyelin. (D) Summary of several key changes across early stages of hyperplasia.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementalData1.xlsx
- SupplementalData2.xlsx
- SupplementalData3.zip
- SupplementalData4.zip
- SupplementalData5.zip
- SupplementalData6.xlsx
- SupplementalData7.xlsx
- SupplementalData8.xlsx
- HTANFAPBulkPaperSupplementsNatureGenetics051121.docx