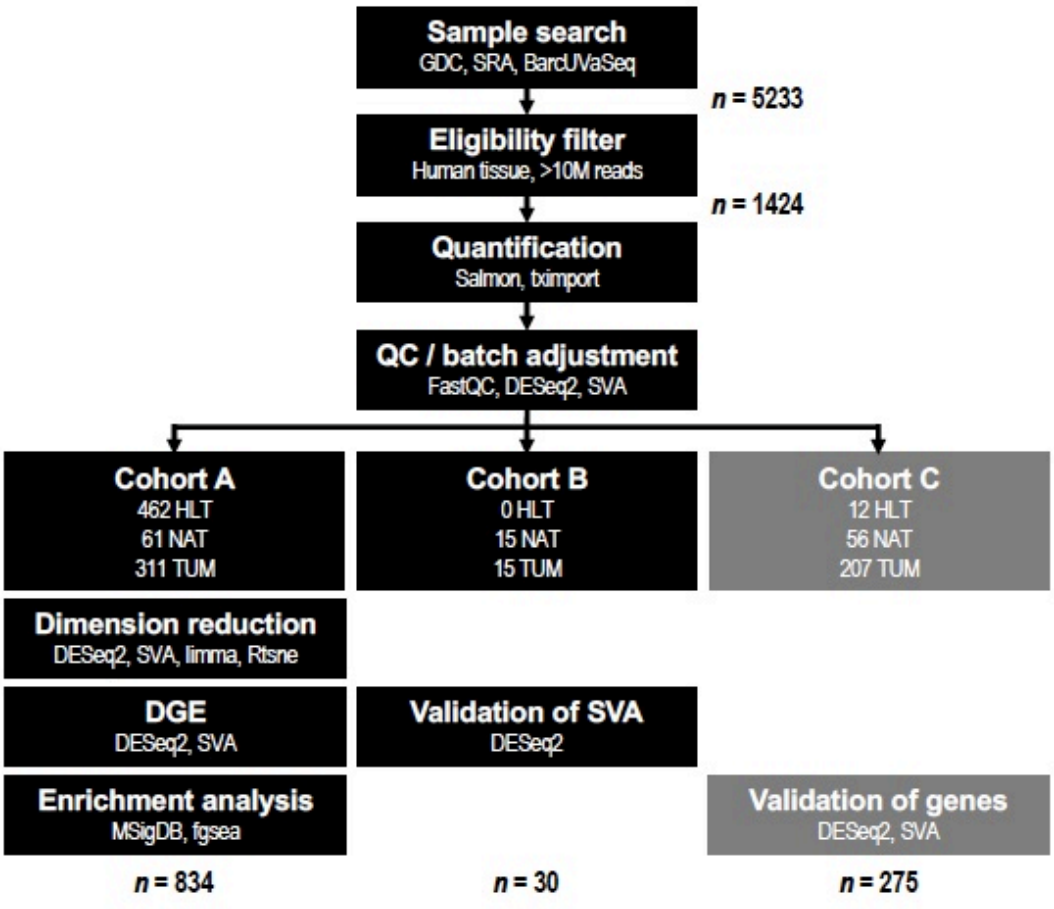
Principal Author's Name: Christopher Dampier
Title of Abstract: Oncogenic features in histologically normal mucosa: novel insights into field effect from a mega-analysis of colorectal transcriptomes

Background: Colorectal cancer is a common malignancy that can be cured when detected early, but recurrence among survivors is a persistent risk. A field effect of cancer in the colon has been reported and could have implications for diagnosis and therapy, but assessments to date have been limited by small sample sizes and inappropriate tissue comparisons. To address these limitations, a joint analysis of pooled transcriptomic data from all available bulk RNA-seq datasets of healthy, histologically normal tumor-adjacent, and tumor tissue was performed.

Methods: Bulk RNA-seq datasets from flash frozen tissue were identified from the Genomic Data Commons, the Sequence Read Archive, as well as from biopsies of non-diseased colon obtained through routine colonoscopy. Primary analyses were limited to samples with a quantified read depth of at least 10 million paired-end reads. Transcript abundance was estimated with Salmon controlling for GC bias, and downstream analysis was performed in R 3.5.1 with the following packages: gene-level counts were aggregated with *tximport*, batch effects were estimated with *SVA*, differential expression was tested with *DESeq2*, and enrichment analysis was performed with *fgsea*.



Results: After sample selection, three study cohorts were created using a total of 1139 colorectal tissue samples. Cohort A, comprising 834 independent samples from eight independent datasets, including 462 healthy, 61 tumor-adjacent, and 311 tumor samples, was the primary analysis cohort used for biological discovery (Table 1). Cohorts B and C were used for validation. Cohort B was composed of paired samples (i.e. tumor and matched normal from the same subject), and Cohort C was composed of single-end sequencing libraries (i.e. as opposed to paired-end).

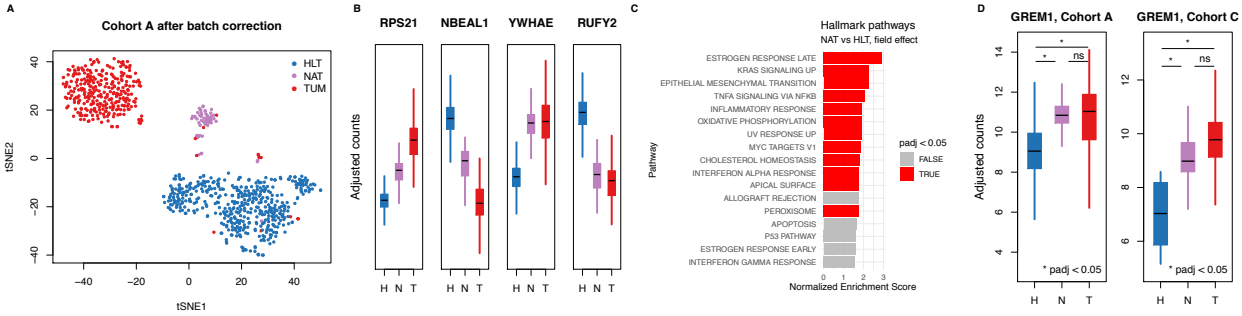
Table 1 Sample counts and demographics for Cohort A

Source	Dataset	# samples				Sex (% female)			Age (mean,sd)		
		HLT	NAT	TUM	Total	HLT	NAT*	TUM*	HLT	NAT*	TUM*
BarcUVa	BarcUVa	260	--	--	260	63%	--	--	60, 7	--	--
GDC	TCGA-COAD	--	36	209	245	--	53%	45%	--	72, 13	65, 13
GDC	TCGA-READ	--	8	81	89	--	88%	46%	--	67, 18	63, 12
SRA	GTEx	202	--	--	202	42%	--	--	49, 13	--	--
SRA	HebeiMU	--	4	5	9	--	75%	60%	--	*NR*	*NR*
SRA	KoreaAMC	--	8	6	14	--	*NR*	*NR*	--	*NR*	*NR*
SRA	KoreaPNU	--	1	2	3	--	0%	100%	--	62, --	73, 4
SRA	Mayo	--	4	8	12	--	50%	38%	--	65, 16	66, 18
Total		462	61	311	834	54%	59%	46%	55, 11	70, 14	65, 13

* missing data on subset of samples

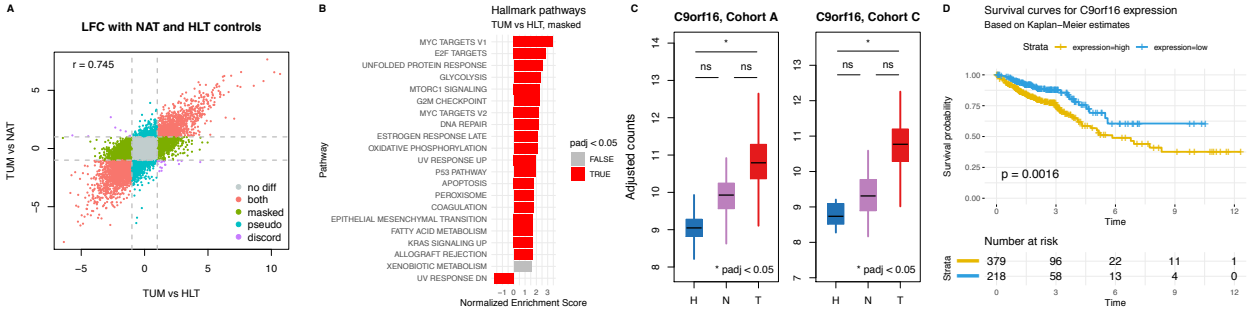
A distinct tumor-adjacent transcriptomic profile was observed (Fig 1A). Tumor-adjacent expression represented an intermediate state between healthy and tumor expression. Among differentially expressed genes in tumor-adjacent samples, 1143 were expressed in patterns similar to tumor samples, and these genes were enriched for cancer-associated pathways (Fig 1B,C). Of the 33 unique genes contributing most to pathway enrichment, 20 were validated in Cohort C (i.e. were found to have the same direction of relative expression between tumor-adjacent and healthy samples in Cohort C as observed in Cohort A). A provocative example was *GREM1* (Fig 1D), which encodes a BMP antagonist ectopically and highly expressed in hereditary mixed polyposis syndrome. Over-expression of *GREM1* in histologically normal tissue would be expected to potentiate malignant transformation.

Fig 1 (A) Dimension reduction analysis showing unsupervised learning on global gene expression. Clusters demonstrate structure in the data. **(B)** Field effect patterns in NAT gene expression. **(C)** Pathways enriched in field effect genes. **(D)** *GREM1* expression levels in discovery and validation cohorts.



Furthermore, 3856 genes were found to differ between healthy and tumor samples that did not differ between tumor-adjacent and tumor samples (Fig 2A). The smaller number of tumor-adjacent samples likely accounted for some of the difference, but the field effect may also have contributed. Most pathways enriched in genes differentially expressed between tumor and control were the same whether healthy samples or tumor-adjacent samples were used as controls (Fig 2B). Although specific genes may have been masked by field effect, co-regulated pathways were generally not. Nevertheless, the availability of true healthy control tissue presented a unique opportunity to discover novel, tumor-associated genes that may otherwise be masked by field effect. A subset of poorly-characterized genes among those differentially expressed between tumor and control specifically when healthy tissue was used as control was identified by filtering for gene symbols including "orf" or beginning with "LOC". Among 62 such genes discovered in Cohort A, 23 were validated in Cohort C, including *C9orf16*, an uncharacterized gene on chromosome 9 that was previously shown to be prognostic among subjects with colorectal cancer (Fig 2 C,D). The tumor-specificity and prognostic value of this relatively unknown gene's expression make it an important target for future investigation.

Fig 2 (A) Scatterplot of transcriptome-wide log fold change between TUM and control samples, where control samples are HLT (x-axis) or NAT samples (y-axis). Colors indicate genes potentially masked (green), misleadingly highlighted (blue-green), or unaffected (red) by field effect. Pearson correlation coefficient is displayed. **(B)** Pathways enriched in genes differentially expressed between TUM and HLT samples but not between TUM and NAT samples. **(C)** *C9orf16* expression levels in discovery and validation cohorts. **(D)** Survival curves for *C9orf16* high and low groups from previously published TCGA data.



Conclusions: Novel insights into the field effect in colorectal cancer were generated in this mega-analysis of the colorectal transcriptome. Oncogenic features that may help explain metachronous lesions in cancer survivors as well as uncharacterized tumor-associated transcripts that may be useful for future diagnosis and treatment were revealed.