scientific reports

Check for updates

OPEN

Transcriptional dynamics of colorectal cancer risk associated variation at 11q23.1 correlate with tuft cell abundance and marker expression in silico

Bradley T. Harris[1], Vidya Rajasekaran[1], James P. Blackmur[1,2], Alan O'Callaghan[2], Kevin Donnelly[1,2], Maria Timofeeva[1,3], Peter G. Vaughan-Shaw[1,2], Farhat V. N. Din[1], Malcolm G. Dunlop[1,2] & Susan M. Farrington[1]✉

Colorectal cancer (CRC) is characterised by heritable risk that is not well understood. Heritable, genetic variation at 11q23.1 is associated with increased colorectal cancer (CRC) risk, demonstrating eQTL effects on 3 cis- and 23 trans-eQTL targets. We sought to determine the relationship between 11q23.1 cis- and trans-eQTL target expression and test for potential cell-specificity. scRNAseq from 32,361 healthy colonic epithelial cells was aggregated and subject to weighted gene co-expression network analysis (WGCNA). One module (blue) included 19 trans-eQTL targets and was correlated with *POU2AF2* expression only. Following unsupervised clustering of single cells, the expression of 19 trans-eQTL targets was greatest and most variable in cluster number 11, which transcriptionally resembled tuft cells. 14 trans-eQTL targets were found to demarcate this cluster, 11 of which were corroborated in a second dataset. Intra-cluster WGCNA and module preservation analysis then identified twelve 11q23.1 trans-eQTL targets to comprise a network that was specific to cluster 11. Finally, linear modelling and differential abundance testing showed 11q23.1 trans-eQTL target expression was predictive of cluster 11 abundance. Our findings suggest 11q23.1 trans-eQTL targets comprise a *POU2AF2*-related network that is likely tuft cell-specific and reduced expression of these genes correlates with reduced tuft cell abundance in silico.

Colorectal cancer (CRC) is the fourth most common cancer type UK and worldwide[1,2]. Approximately 40% of CRC risk is attributable to heritable genetic variation[3], with rare, highly penetrant mutations responsible for a small fraction of overall risk[4,5]. Genome wide association studies (GWAS) have identified 129 common genetic variants associated with CRC risk[6,7]. Several common CRC genetic risk variants are associated with heritable changes in the expression levels of genes in colonic mucosa, known as expression quantitative trait loci (eQTLs)[8–11].

Genetic variation at 11q23.1 is associated with increased CRC risk[12]. However, the large number of variants in high linkage disequilibrium at 11q23.1 makes identifying the causal variant, a key step to identifying the mechanism of gene dysregulation, difficult. Studies have shown that CRC risk associated variation at several 11q23.1 variants is correlated with the downregulation of three local genes; *POU2AF2* (*also known as C11orf53*), *COLCA1, COLCA2*[10,13,14], known as cis-eQTL targets. We have recently shown that variation at rs3087967, a single nucleotide variant in the 3′ UTR of *POU2AF2*, is correlated with distal CRC risk and a myriad of distant, trans-eQTL targets throughout the colon[11], Supplementary Data 1. Of these, only two have a common, well-described function; *IL17RB* and *TRPM5* are experimentally determined markers of tuft cells—a rare epithelial cell-type[15,16]. The function of several other trans-eQTL targets is currently unknown and their exact relevance to both 11q23.1 cis-eQTL target expression and CRC risk is not characterised.

[1]Edinburgh Cancer Research, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [2]MRC Human Genetics Unit, Institute of Genetics and Cancer, University of Edinburgh, Edinburgh, UK. [3]Department of Public Health, D-IAS, Danish Institute for Advanced Study, University of Southern Denmark, Odense, Denmark. ✉email: susan.farrington@ed.ac.uk

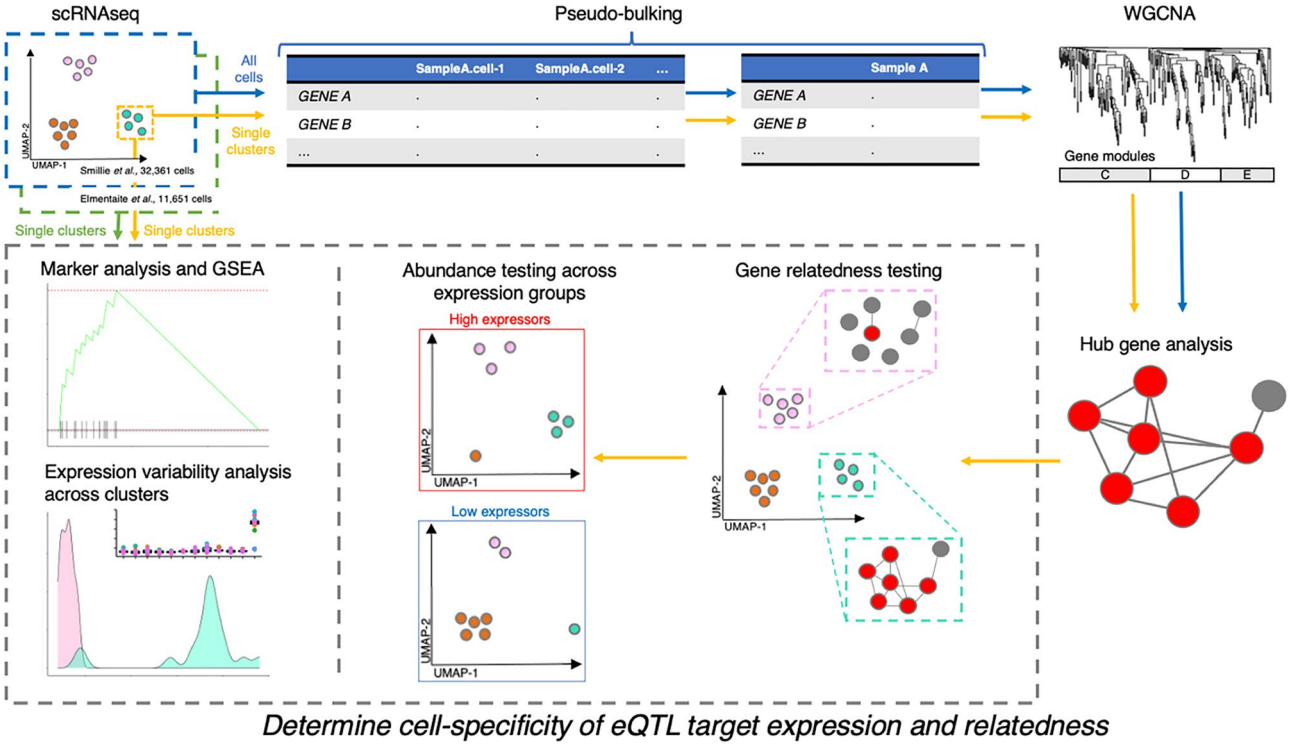

Figure 1. Overview of study design and analysis. This study makes use of two single cell RNA sequencing (scRNAseq) datasets: Smillie et al.[22]—n = 32,261 and Elmentaite et al.[23]—n = 11,651. The analysis performed on each dataset is outlined by arrow colour: blue = all cells from Smillie et al.[22], yellow = individual clusters from Smillie et al.[22], green = individual clusters from Elmentaite et al.[23]. *WGCNA* weighted gene co-expression network analysis, *GSEA* gene set enrichment analysis.

Current methods of eQTL detection, while widely used, exhibit some key limitations. eQTLs are frequently identified by linear models of bulk RNA-seq/microarray transcriptome analysis of healthy tissue[17–19]. These methods often treat both gene expression and single-nucleotide polymorphisms as independent, linear entities: assumptions which over-simplify the complex relationships governing gene expression dynamics and limit results to additive gene-dosage related findings. In addition, eQTL analyses require performance of an extremely large number of independent tests to be conducted, limiting its sensitivity. Correlation-based methods of gene expression analysis, such as weighted gene co-expression network analysis (WGCNA)[20], circumvent this problem by agnostically identifying correlations between individual genes, and entire, non-overlapping gene modules, with binarized categorical or quantitative traits. Modules of correlated genes may themselves be correlated with sample phenotypes. In addition, WGCNA also does not require a hard thresholding of correlations, a major advantage over other correlation-based methods that rely on arbitrary cut offs. We have recently shown that WGCNA can be efficacious at identifying genes driving transcriptional dynamic changes in the colorectum of patients treated with Vitamin D[21], even in the absence of statistically significant changes in differential expression analysis. Moreover, as CRC risk-associated eQTL targets are identified by bulk expression methods, findings are inherently limited in their potential to detect cell-specific changes, particularly in rare cell-types; expression changes within which may be masked due to low relative abundance.

We hypothesised that the expression of 11q23.1 trans-eQTL targets may be correlated with only a single cis-eQTL target, a relationship which may in turn be specific to a single epithelial cell-type in the colon. We utilised WGCNA across and within single-cell RNA sequencing (scRNAseq) clusters to further characterise eQTL target expression relatedness in colonic epithelial cell-types. An overview of the analysis performed in this study is shown in Fig. 1.

Results

Understanding cis-eQTL specificity of 11q23.1 trans-eQTL effects. To test potential colonic epithelial cell-specificity of 11q23.1 eQTL effects, we sought to analyse the expression of target genes within colonic epithelial cell-types. To this end, we obtained scRNAseq from 32,361 healthy human colonic mucosa epithelial cells from eleven individuals[22]. To first assess the validity of this dataset in studying 11q23.1 variation-related expression dynamics, we devised a method of pseudo-bulking the expression from all cells in this dataset, aggregating the expression of every gene across all cells from each sample (see Fig. 1, "Methods" section). Pseudo-bulked expression from 11q23.1 nominally significant trans-eQTL targets, present in the scRNAseq dataset ($p < 0.01$, $n = 273$), Fig. 2a, was then subject to WGCNA[20]. Genes were agnostically grouped into modules of correlated expression, and the correlation between the eigenvector (first principal component) of each

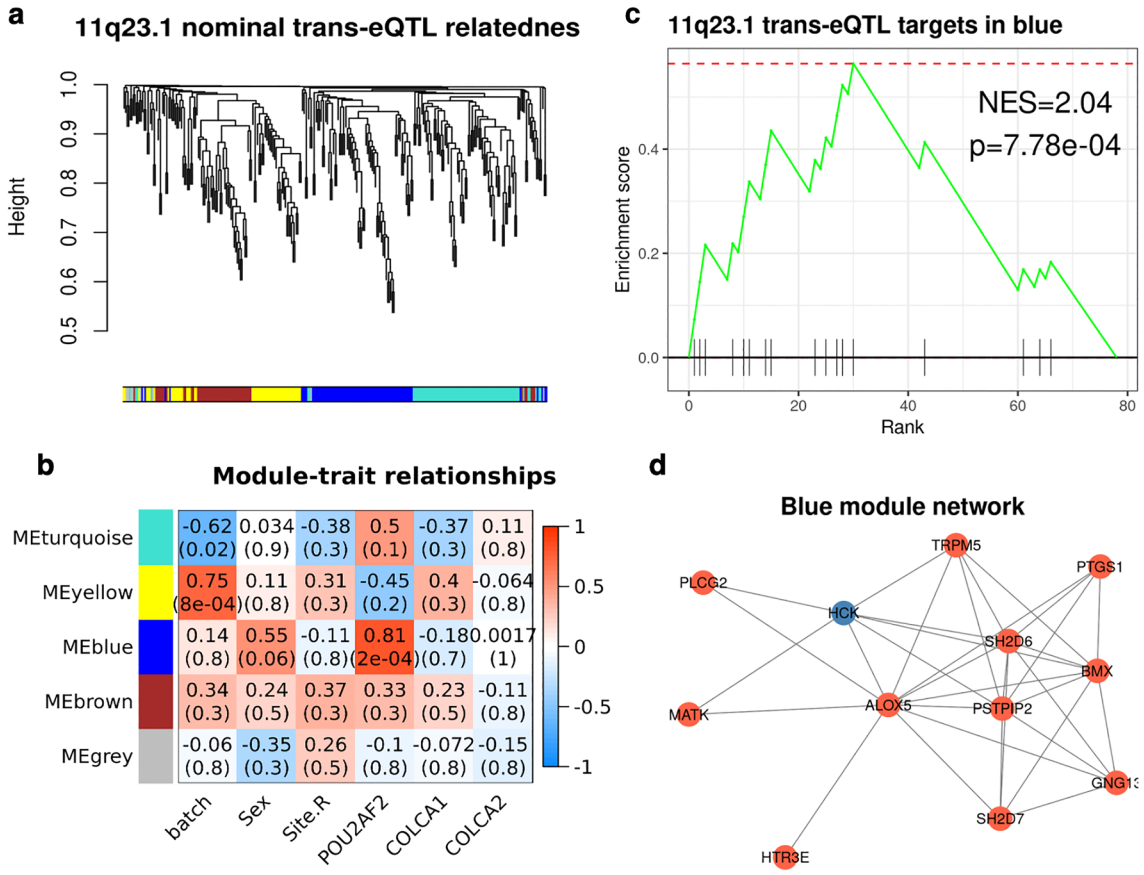

Figure 2. 11q23.1 trans-eQTL targets are correlated with the expression of *POU2AF2,* but not *COLCA1* or *COLCA2* in pseudo-bulked scRNASeq. (**a**) Hierachical clustering of complete distances between pairwise correlations of 11q23.1 nominal trans-eQTL targets pseudo-bulked expression ($p < 0.01$, $n = 273$) in 32,361 healthy colonic epithelial scRNAseq[22]. (**b**) Weighted Gene Co-expression Network Analysis (WGCNA[20], identified module trait relationships. Pearson correlations shown above Benjamini-Hochberg[45] corrected p-values in brackets. (**c**) Gene Set Enrichment Analysis (GSEA) of 11q23.1 trans-eQTL targets (FDR < 0.05) in blue module genes, ranked by their module membership. (**d**) Kamadakawai network of Blue module relatedness (adjacency > 0.3). Red nodes indicate FDR < 0.05 trans-eQTL target of 11q23.1.

module with sample traits and pseudo-bulked cis-eQTL target gene expression was calculated, Fig. 2b. The blue module, comprising 77 genes, was found to correlate with *POU2AF2* expression (cor = 0.81, FDR = 2e−04), but none of the sample traits, *COLCA1* or *COLCA2* expression. The blue gene module includes 17 of the 20 significant 11q23.1 trans-eQTL targets that passed gene filtration quality control (FDR < 0.05, Vaughan-Shaw et al.[11]); *ALOX5, SH2D6, TRPM5, BMX, PSTPIP2, GNG13, IL17RB, HTR3E, PTGS1, SH2D7, OGDHL, MATK, PLCG2, LRMP, PIK3CG, HTR3C* and *CAMP*, thus indicating the genes comprising this module are specifically related to the expression of *POU2AF2,* Supplementary Table S2.

To assess the contribution of 11q23.1 genes to the correlation of the blue module with *POU2AF2*, we performed gene set enrichment analysis of the 11q23.1 trans-eQTL target genes (FDR < 0.05) against the genes in the blue module, ranked by their module membership—a measure of their relatedness to all other genes in the module[20]. We found the genes in this module are highly enriched for 11q23.1 trans-eQTL targets, normalised enrichment score (NES) = 2.04, p = 7.78e−04, Fig. 2c. In addition, 11 trans-eQTL targets comprised the 12 genes with the greatest adjacency in the blue module (adjacency > 0.3), Fig. 2d. We also find that module membership is highly correlated with gene significance for *POU2AF2* in the blue module, reinforcing the overall correlation of the network with *POU2AF2* expression (Supplementary Fig. S1). Together, this replicates the trans-eQTL target expression relatedness previously described[11] and indicates that the expression of the majority of significant 11q23.1 trans-eQTL target genes is correlated with *POU2AF2* specifically in this dataset.

Analysing cell-specificity of 11q23.1 eQTL target expression. To test the potential for cell-specific expression of 11q23.1 eQTL targets in this dataset, we performed dimensionality reduction and clustering of the single cells, identifying a total of 12 transcriptionally distinct cell-clusters, named '0'–'11', Fig. 3a. We then calculated the markers of each cluster and found the markers of cluster 11, comprising 318 cells, includes fourteen 11q23.1 trans-eQTL targets (FDR < 0.05), Table 1 (full list of cluster 11 marker genes available in Supplementary Data 2). In addition, cluster 11 markers are significantly enriched for 11q23.1 trans-eQTL targets (NES = 2.15, p = 7.52e−06), Fig. 3b. Cluster 11 was found to transcriptionally resemble the Smillie et al.[22], tuft cell cluster,

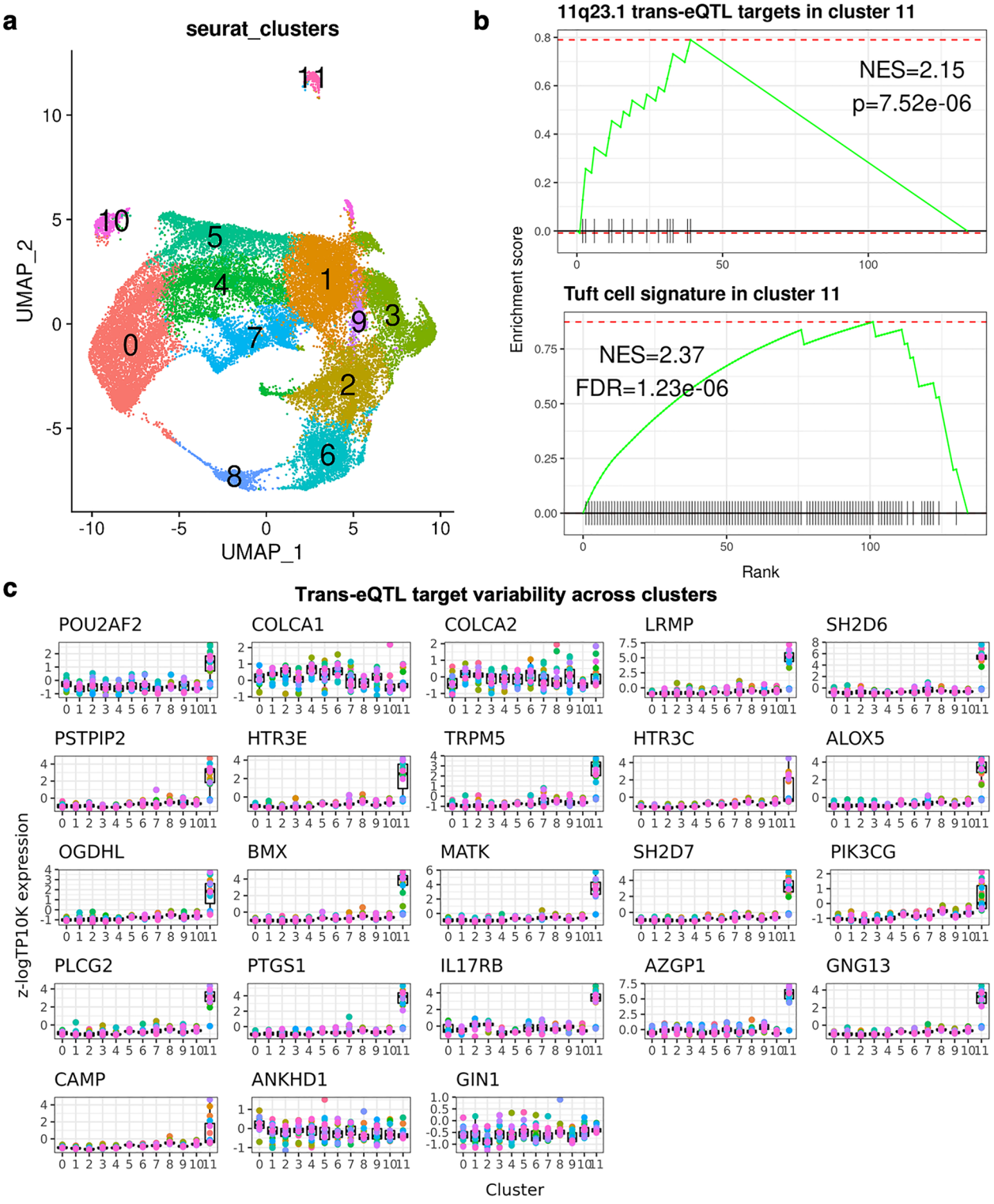

Figure 3. 11q23.1 trans-eQTL expression demarcates tuft-like cell cluster. (**a**) UMAP of 32,361 epithelial scRNASeq, coloured by cell cluster—identified using Seurat[42]. (**b**) Upper: GSEA of 11q23.1 trans-eQTL targets[11]; FDR < 0.05) in cluster 11 markers. Lower: GSEA of putative colonic tuft cell signature[22] in cluster 11 markers. *p* p value, *FDR* false discovery rate. (**c**) Relative, pseudo-bulked expression (log transcripts per 10,000) of 11q23.1 trans-eQTL targets (FDR < 0.05) across clusters.

Gene	p_val	avg_log2FC	pct.1	pct.2	p_val_adj
AZGP1	1.00E−279	2.5735445	0.72	0.029	1.00E−279
SH2D6	1.00E−279	2.43533298	0.736	0.007	1.00E−279
LRMP	1.00E−279	1.97284472	0.572	0.005	1.00E−279
PTGS1	1.18E−278	1.33766533	0.459	0.003	1.75E−274
IL17RB	1.00E−264	1.04046297	0.412	0.031	1.48E−260
BMX	7.18E−248	1.18575976	0.406	0.002	1.07E−243
ALOX5	8.25E−244	1.09872574	0.409	0.003	1.22E−239
MATK	2.88E−240	1.3710223	0.406	0.003	4.28E−236
SH2D7	2.50E−235	1.20590753	0.393	0.002	3.72E−231
GNG13	2.93E−220	0.98868335	0.362	0.001	4.34E−216
TRPM5	1.32E−214	0.94972732	0.377	0.004	1.96E−210
PLCG2	1.22E−209	0.98215989	0.365	0.002	1.82E−205
PSTPIP2	2.21E−197	0.87613246	0.333	0.004	3.28E−193
HTR3E	4.86E−184	0.8684383	0.308	0.001	7.21E−180

Table 1. 11q23.1 trans-eQTL targets identified as cluster 11 markers. Markers calculated using MAST[49]. *Avg_log2FC* average log2 fold change between cluster 11 and all other clusters, *Pct1* proportion of gene expression in cluster 11, *Pct2* proportion of expression in non-cluster 11, *p_val_adj* FDR-corrected p-value.

by enrichment of each cluster's markers with the authors' putative markers (NES = 2.37, FDR = 1.23e−06, see "Methods" section). Together, this indicates cluster 11 is transcriptionally defined by the expression of 11q23.1 trans-eQTL targets, several of which are themselves putative tuft cell markers.

Because the expression of 11q23.1 trans-eQTL targets is reduced in individuals with the CRC risk-associated genotype at 11q23.1[11], we wanted to assess the variability in the expression of these genes within each cluster. We pseudo-bulked the expression of all cells from each sample, within clusters, and analysed 11q23.1 trans-eQTL target expression (Fig. 3c). The relative expression level and variability of *POU2AF2* and 18 of the 11q23.1 trans-eQTL targets is overwhelmingly greatest within cluster 11, indicating the eQTL effect on these genes is exacerbated within this cluster. Notably, the relative variation and expression of cis-eQTL targets *COLCA1* and *COLCA2* and trans-eQTL targets *ANKHD1* and *GIN1,* is not greatest within this cluster, suggesting the eQTL effect on these genes may not be driven by transcriptional dynamics within this cell-type. In addition, we analysed 11q23.1 eQTL target variability at the single cell level, Supplementary Fig. S2. The variability of *POU2AF2* and the same 18 trans-eQTL targets was found to be greatest within cluster 11, replicating the potential exacerbation of the eQTL effect within this cluster and supporting the validity of our pseudo-bulk method. *POU2AF2* expression, however, was not found to demarcate cluster 11 by marker identification analysis.

To test the robustness of tuft cell-like mapping of 11q23.1 trans-eQTL target expression and variability, we replicated this analysis in an independent dataset of 11,651 healthy adult colonic epithelial cells from 3 individuals[23], Supplementary Fig. S3. In this instance, we identified 19 cell-clusters by dimensionality reduction and unsupervised clustering, named '0'–'18'. The markers of cluster 18 were significantly enriched for the expression of 11q23.1 trans-eQTL targets (NES = 2.50, p = 5.52e−09), 11 of which were identified as markers of this cluster, Supplementary Table S3. Cluster 18 was also enriched for the tuft cell signature (NES = 2.41, p = 7e−08), replicating our previous finding. The relative variability and expression of the majority of 11q23.1 trans-eQTL targets was also greatest within cluster 18. The expression of 13 trans-eQTL targets was also most variable within cluster 18 when expression was analysed at the single cell level, Supplementary Fig. S4. Interestingly, *POU2AF2* expression variability was not found to vary greatest within cluster 18 when using expression from single cells.

Understanding 11q23.1 cis- and trans-eQTL relatedness within clusters. The demarcation and variability of 11q23.1 trans-eQTL target expression within a tuft-like cell cluster strongly indicates the eQTL effect is derived from altered gene expression in this cell-type specifically. However, it is possible that the gene–gene relatedness of 11q23.1 eQTL targets, including those with *POU2AF2*, is not specific, but rather exacerbated in this cluster. To test this, we sought to divide samples by genotype at rs3087967, the variant associated with expression changes of trans-eQTL targets[11], and analyse the congruence of trans-eQTL target expression relatedness within clusters. While genotype information was unavailable for the Smillie et al.[22], dataset, raw sequencing reads were available for the Elmentaite et al.[23] dataset and as rs3087967 lies within the 3′ UTR of *POU2AF2*, we performed variant calling on these samples using orthogonal tools, Supplementary Table S4. Using freebayes[24], we found that all samples were called as homozygous for the non-risk allele at rs3087967, except for one sample called as heterozygous. However, as all 3 other samples from this individual were called as homozygous non-risk, it is likely this is a technical error. Using bcftools[25], all samples were identified to be homozygous non-risk for rs3087967. The lack of genetic variation at rs3087967 in these samples is consistent with the absence of heightened *POU2AF2* variability in cluster 18 and indicates this dataset would be unlikely to be of use to identify *POU2AF2*-related expression dynamics. The relatively high variability of trans-eQTL target expression within cluster 18 may be indicative of non-11q23.1-related dynamics, such as changes during differentiation or cell cycle progression.

To test the potential efficacy of the Smillie et al.[22], dataset to further study the 11q23.1 eQTL target expression dynamics, we compared the standardized variability of 11q23.1 trans-eQTL targets within the respective demarcated clusters at the single-cell level, Supplementary Table S5. We found that the expression variability of 14 of the 15 11q23.1 trans-eQTL targets, expressed in both Elmentaite et al.[23], cluster 18 and Smillie et al.[22], cluster 11, was significantly increased in the latter (fold-change range 1.46–9.7, median = 1.97, mean = 2.83, 95% confidence interval = 1.50–4.18, 100,000 permutation p < 1e−5). The only 11q23.1 trans-eQTL target not to exhibit increased variability in Smillie et al.[22], cluster 11 was OGDHL (fold change = 0.73). Notably, the expression of *POU2AF2,* was also greater in Smillie et al.[22], cluster 11 (fold change = 1.94). Subsequent analysis was therefore focussed on the Smillie et al.[22], dataset.

To divide samples into high and low expressors of *POU2AF2*-expression related genes, we hierarchically clustered samples based on the relative expression of pseudo-bulk WGCNA-identified blue module hub genes. Hub genes were defined by module membership > 0.7, intramodular connectivity > 0.7 and network adjacency > 0.3 and include 10 genes: *TRPM5, PSTPIP2, SH2D6, ALOX5, BMX, GNG13, SH2D7, HCK, PLCG2, MATK*, Fig. 4a. Five samples were separated at the first branch of clustering and exhibit a strong relative reduction in the expression of these genes. This grouping of samples is henceforth referred to as the 'blue module hub gene grouping'. To assess the significance of this separation at representing underlying transcriptional differences, we performed 10,000 permutations, using 10 randomly sampled genes, and generated a p-value of 0.055.

To first test gene–gene relatedness within cluster 11, we performed WGCNA on the relative, pseudo-bulked expression of the top 5000 most variable genes within this cluster, identifying a total of 20 modules, 7 of which exhibited correlations with sample covariates that approached significance (FDR < 0.1), Fig. 4b. We found one module, 'cluster 11 black', which was highly correlated with both the 'blue hub gene grouping' (cor = 0.72, FDR = 0.032) and *POU2AF2* expression (cor = 0.68, FDR = 0.048). 'Cluster 11 black' is comprised of 290 genes, including fifteen 11q23.1 trans-eQTL targets, Supplementary Data 3. As the 'blue module hub gene' grouping is derived from analysis of genes correlated with *POU2AF2* across all cells, the correlation of 'cluster 11 black' with both this grouping and *POU2AF2* expression indicates that this relationship is preserved within and potentially derived from this cell-cluster.

We then sought to test whether the gene–gene relatedness of 11q23.1 trans-eQTL targets, correlated with *POU2AF2*, was specific to cluster 11. To this end, we defined an auxiliary module, comprised of the twelve 11q23.1 trans-eQTL targets correlated with *POU2AF2* (cor > 0.5, p < 0.05) in cluster 11 black, which exhibited high module membership (MM.black > 0.5), Fig. 4c. These genes include: *HTR3E, LRMP, GNG13, ALOX5, SH2D7, PTGS1, MATK, BMX, AZGP1, IL17RB, SH2D6* and *OGDHL*. We performed pairwise module preservation of this auxiliary module, using identical parameters as used for intra-cluster 11 analysis, in all other clusters, see "Methods" section. For two modules, cluster 8 and cluster 10, the 5000 most variable genes included only a single member of the auxiliary module and so were excluded from this analysis. To analyse the preservation of the connectivity of these genes, we assessed the similarity of each gene correlation with the module eigengene, and the equivalent values in cluster 11 (cor.kME), Fig. 4d. No module exhibited a significant cor.kME with cluster 11 (p > 0.05), indicating there is low overall preservation of the connectivity between these genes in all other clusters. The average module membership of genes in this module (average.MM) was also reduced in all clusters compared with cluster 11, Fig. 4e. Finally, we analysed the pairwise gene–gene correlations between all members of the auxiliary module and *POU2AF2* within each cluster, Fig. 4f. While there are rare correlations between these genes in other modules (cor > 0.5, p < 0.05), all comparisons reach this threshold in cluster 11. Together, this evidence indicates these 12 trans-eQTL targets comprise a transcriptional network that is correlated with *POU2AF2* expression and likely specific to cluster 11.

Identification of cluster 11 abundance associated genes. As many of the 11q23.1 eQTL targets, including those comprising the cluster 11-specific network, demarcate this cluster, we wanted to examine the relationship between their expression and cluster 11 abundance. First, we performed linear modelling of the relative abundance of cluster 11 and the pseudo-bulked expression of *POU2AF2* only. We found that the expression of *POU2AF2* is associated with the relative abundance of cluster 11 (coefficient = 0.389, p = 0.00431) indicating potential for the expression of *POU2AF2* alone to be moderately predictive of the abundance of this cell-type, Fig. 5a.

To agnostically test the predictive power of 11q23.1 trans-eQTL target expression on cluster 11 abundance, we tested the association between the expression of all genes and the proportion of cluster 11 in samples, Fig. 5b. We found that all genes significantly associated with the abundance of this cluster (FDR < 0.05, log-fold change > 1) were indeed 11q23.1 trans-eQTL targets, including: *ALOX5, BMX, GNG13, MATK, SH2D7, PSTPIP2, TRPM5* and *PTGS1*. In fact, the strength of gene association with cluster 11 abundance was also significantly enriched for 11q23.1 trans-eQTL targets (NES = 2.15, p = 8.03e−10, Fig. 5c).

While our linear modelling strongly supports a predictive role of 11q23.1 eQTL target expression in the abundance of cluster 11, we wanted to agnostically test whether the expression of the *POU2AF2*-related trans-eQTL targets was correlated with abundance changes to any cluster. To this end, we utilised miloR[26] to calculate cell-neighbourhoods, Supplementary Fig. S5, which were then used to perform differential abundance testing across the 'blue module hub gene grouping', defined in Fig. 4a. To generalise neighbourhoods to the cell clusters we have identified, we subsequently filtered for neighbourhoods which represented a majority of a single cluster (majority proportion > 0.8), Fig. 5d. No neighbourhood from cluster 10 passed this threshold and so this cluster was excluded. In the low blue hub gene group, there are significant reductions (Spatial FDR < 0.01) from cluster 0 and increases in neighbourhoods in clusters 1, Fig. 5d. The change in abundance of these neighbourhoods represent a small proportion of the total number of neighbourhoods detected in these clusters (2.1% and 1.3% respectively) and are therefore unlikely to represent a dramatic phenotype. In contrast, all 7 of the

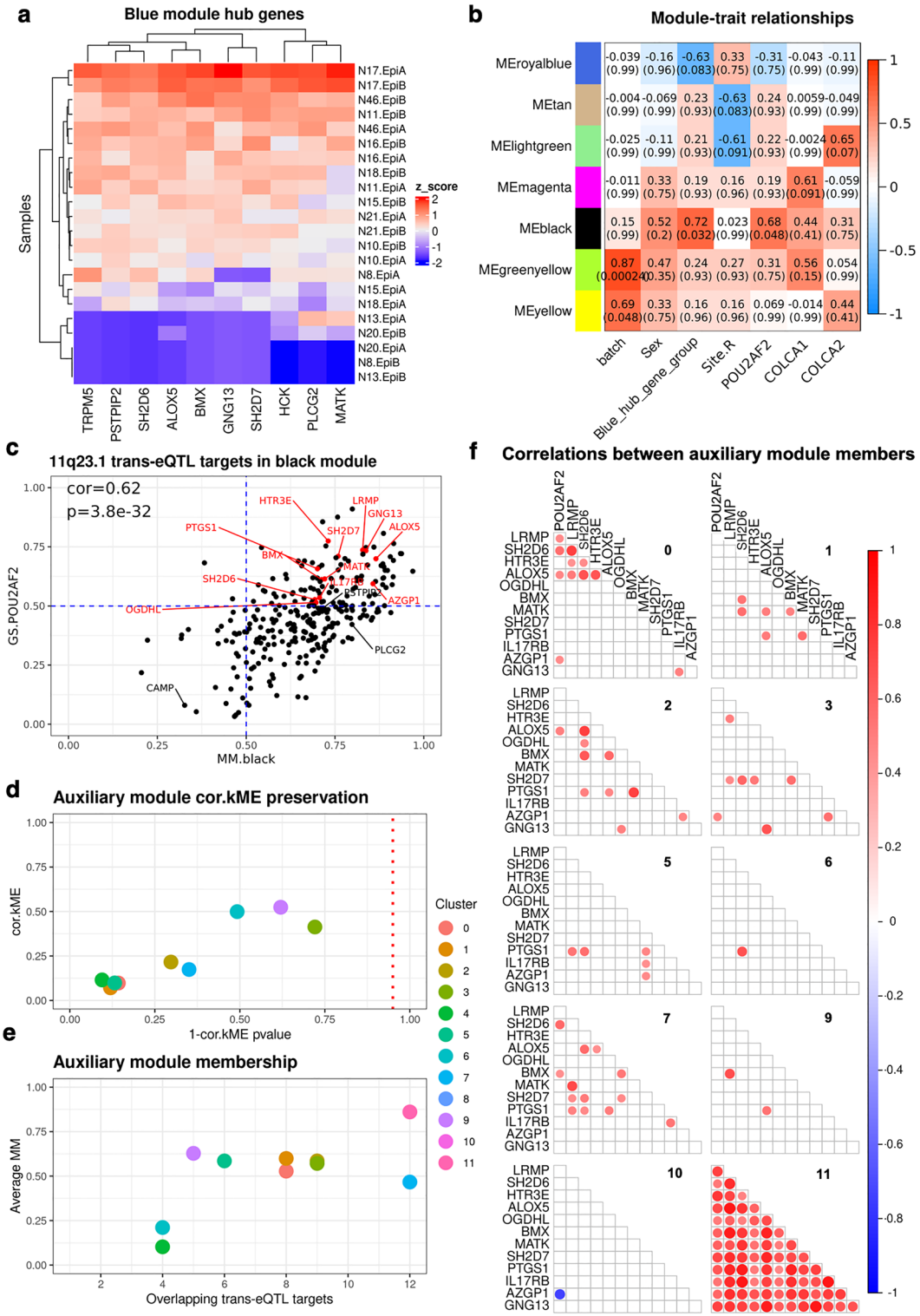

Figure 4. Several 11q23.1 trans-eQTL targets comprise *POU2AF2*-correlated network in tuft-like cluster only. (**a**) Relative (z-score) expression of blue module hub genes from Fig. 1d (MM > 0.7, kIM > 0.7, adj > 0.3). Robustness assessed by 10,000 permutations: p = 0.055. (**b**) Module trait matrix of gene modules identified by WGCNA within pseudo-bulked expression from cluster 11. Only modules exhibiting a correlation with a covariate (FDR < 0.1) are shown. Total number of modules = 20. (**c**) Correlation of Gene Significance (GS) and Module Membership (MM) of genes in the black module from (**b**). 11q23.1 trans-eQTL targets are highlighted. 11q23.1 trans-eQTL targets with GS.POU2AF2 > 0.5 and MM.black > 0.5 (red) were used as the auxiliary module. (**d**) Preservation of correlations between auxiliary module genes with the module eigengene (ME) and equivalent values in cluster 11. Dashed line indicates nominal significance threshold (p = 0.05). (**e**) Average MM of genes within the auxiliary module across clusters. (**f**) Pairwise Pearson correlations (p < 0.05) between the pseudo-bulked expression of auxiliary module genes across clusters. Cluster number in upper right of each facet. Cluster 4 and 8 exhibited no significant (p < 0.05) correlations and so are not plotted.

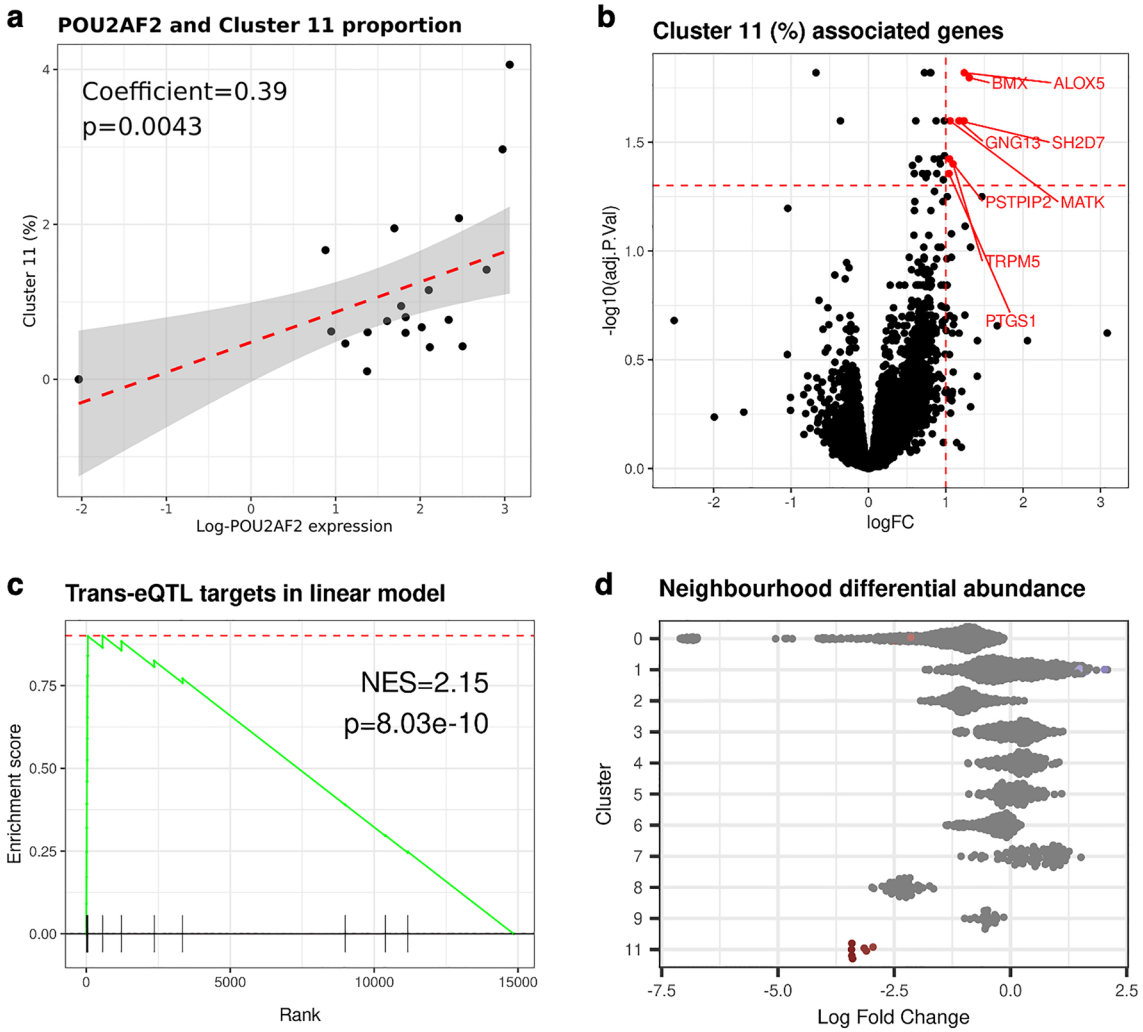

Figure 5. 11q23.1 trans-eQTL target expression is associated with the abundance of tuft cell-like cluster. (**a**) Linear modelling of pseudo-bulked *POU2AF2* expression and cluster 11 abundance. (**b**) Volcano plot of linear model result on pseudo-bulked expression of all 14,843 genes. Significantly associated 11q23.1 trans-eQTLs (logFC > 1, FDR < 0.05) highlighted. (**c**) GSEA of 11q23.1 trans-eQTLs (FDR < 0.05) in (**b**), ranked by logFC. (**d**) Differential abundance of neighbourhoods in the 'Low' blue hub gene group compared to 'High'. Neighbourhoods were identified using miloR[26] and grouped by cluster. Only neighbourhoods with major cluster proportion > 0.8 are plotted and only significant neighbourhoods (Spatial FDR < 0.01) are coloured by their logFC (red = down, blue = up). No neighbourhoods from cluster 10 were comprised of major proportion > 0.8.

neighbourhoods comprising a majority of cluster 11 cells were significantly underrepresented in the low 'blue module hub gene group'. These results indicate 11q23.1 eQTL target expression is correlated with a considerable and likely specific change in abundance of cluster 11 cells.

Discussion

In this study, our pan-cluster WGCNA serves as validation of the expression relatedness between trans-eQTL targets previously identified to be correlated with CRC associated variation at 11q23.1[11]. 11q23.1 trans-eQTL target expression was also found to be more correlated with *POU2AF2* over other cis-eQTL targets. Following clustering of single cells, many of the genes demarcating a single cluster, number 11, were found to be 11q23.1 trans-eQTL targets. Enrichment of these markers against putative gene sets showed cluster 11 transcriptionally resembled tuft cells, replicated in an independent dataset. WGCNA within cluster 11 identified several 11q23.1 trans-eQTL targets that exhibited a high level of relatedness, subsequent analysis of the preservation of this relatedness indicated this was likely to be specific to this cell cluster. Finally, samples which exhibited low overall expression of 11q23.1 trans-eQTL targets most correlated with one another, were found to exhibit a specific and dramatic reduction of the tuft cell-like cluster. Therefore, our results potentiate that transcriptional dynamics correlated with CRC risk associated variation at 11q23.1, are indicative of a reduced abundance of colonic tuft cells.

To our knowledge, this is the first study to map CRC risk associated eQTL targets to specific epithelial cell-types. Heritable inflammatory bowel disease risk loci have recently been associated with shifts in transcriptional

dynamics in individual colonic epithelial cells[22] and it is possible that other CRC risk variants, with robust eQTL effects, are associated with cell-specific changes in transcriptional dynamics. Delineating the cell-specific expression of CRC risk-associated eQTLs may provide valuable insights into the mechanism of risk-associated pathophysiology and should be a focus of future work. The growing size and availability of scRNAseq datasets will likely make cell-specificity mapping of heritable disease associated eQTLs significantly easier, particularly if genotype data is available. Indeed, our study is also not the first use of WGCNA to detect gene–gene relatedness in scRNAseq data. WGCNA has been utilised to identify gene modules associated with the activation of neuronal stem cells[27] and human induced pluripotent stem cells[28], however, like our own study, these studies did not utilise expression from individual cells as the input for WGCNA.

Surprisingly, the expression of the vast majority of 11q23.1 eQTL targets mapped to the cell-type that transcriptionally resembles tuft cells, including; *LRMP, IL17RB, SH2D6, PLCG2, PSTPIP2, TRPM5, SH2D7, AXGP1, PTGS1, ALOX5, BMX*. Many of the most significant 11q23.1 trans-eQTL targets, such as *LRMP, SH2D7* and *ALOX5*, have not previously been associated with specific expression in this cell-type, potentiating their status as a marker in the colonic epithelium. Additional markers of the tuft cell-like cluster include *HCK* and *HPGDS,* for which there is some orthogonal evidence of expression within tuft cells[29,30]. This improves our confidence that cluster 11 represents this cell-type and is not an artefact of the analysis.

Notably, our pan- and intra-cluster WGCNA indicates decoupling of 11q23.1 cis-eQTL targets, suggesting trans-eQTL target expression is attributed to the expression of *POU2AF2* but not *COLCA1* or *COLCA2*. While the Smillie et al.[22], dataset is not genotyped, numerous studies identify *POU2AF2*, *COLCA1* and *COLCA2* as eQTL targets of 11q23.1 variation[10,11,13,14], supporting our use of their expression as a proxy of 11q23.1 genetic variation. In addition, the 11q23.1 cis-eQTL targets are highly correlated with one another in the bulk expression-based studies, our observation of divergence in their expression across transcriptionally distinct cell clusters is consequently novel. Delineation of the 11q23.1 transcriptional dynamics toward *POU2AF2* specifically, implies the association between *POU2AF2* expression and tuft cell abundance is a potentially causal feature of CRC risk. However, as these findings are based on correlation-based analyses in silico, causal relationships are only inferred. Experimental testing of such using gene knock out models is required to both confirm *POU2AF2* potential causality and assess whether *COLCA1* or *COLCA2* confer causality.

Recent studies have identified a direct interaction between POU2AF2 and master transcriptional regulator of the tuft cell lineage, POU2F3[31,32]. These studies indicate that in cell line models of a tuft cell-like subtype of small cell lung cancer, POU2AF2 acts as a transcriptional coactivator of POU2F3 targets, including 11q23.1 trans-eQTL targets *PTGS1*[31] and *AVIL*[32]. While *POU2F3* was not identified as an initial 11q23.1 trans-eQTL target, it was found to correlate with *POU2AF2* expression in cluster 11, Supplementary Table S3. As the 11q23.1 trans-eQTL targets we find correlated with *POU2AF2* putatively demarcate tuft cells in our analyses, direct interaction with POU2F3 is a potential mechanism by which POU2AF2 mediates their expression along with tuft cell differentiation and determination in the colon. Interestingly, *POU2AF2* expression was also found to be positively correlated with small cell lung cancer cell survival in vitro and in vivo[32]. While we find reduced *POU2AF2* expression to be associated with CRC risk, the functional interaction between POU2F3 and POU2AF2, in association with 11q23.1 eQTL target expression, is concordant with the cluster-specific transcriptional dynamics we observe.

It is noteworthy that many of the genes with expression found to be in association with *POU2AF2* in the scRNAseq data were indeed identified as trans-eQTL targets in the bulk analysis[11]. While the mapping of the expression of these genes to a tuft-like cell cluster could only be achieved by using expression from individual cells, the prior identification of these genes by bulk analysis is a testament to the power of such approaches, and concordance of their findings with single cell-based methods.

Finally, the overall potentiation of tuft cell perturbation is of great interest regarding characterisation of the mechanism governing CRC risk at 11q23.1. Tuft cells are associated with stem-, neurotransmitting- and immune-related functions[33–36], however much of the evidence regarding their function is derived from other organs and cannot necessarily be extrapolated to the colon. Interestingly, genetic ablation of tuft cell abundance has been associated with exacerbated tumour progression in mouse models of pancreatic cancer[37,38]. Both studies show this is likely in association with perturbed immune cell function and signalling. In line with this, tuft cell abundance has recently been shown to be reduced in quiescent ulcerative colitis patients[39], suggesting tuft cells are involved in immune regulation in the colon. Future work should aim to experimentally validate the relationship between 11q23.1 variation and tuft cell abundance, examine how this impacts tumourigenesis and identify the potential mechanism of CRC risk predisposition.

Methods

Pre-processing, dimensionality reduction and clustering of scRNAseq data. In the analysis of Smillie et al.[22], scRNAseq data, samples from one patient, N51, were removed as they were found to be outliers on the basis of cell-level mitochondrial and ribosomal protein gene expression, in addition to principal component analysis of pseudo-bulked expression. Fastq files of the Elmentaite et al.[23], scRNAseq data were aligned to the hg19 transcriptome using the 10× Genomics Cell Ranger v3.02 pipeline[40] to produce raw gene-level counts.

All subsequent expression analysis was completed in R version 4.0.2. Once raw counts had been obtained for both datasets, bad quality droplets were filtered by a series of quality control steps: (i) potential empty droplets were detected by thresholding at the inflection point of the barcode rank plot of cells, calculated using *DropletUtils* v1.8.0[41], (ii) genes expressed in fewer than 20 cells were removed, (iii) cells with an expression sparsity > 0.99 were removed, (iv) cells with proportion of mitochondrial gene expression greater than 2.5× (median absolute deviation) from the median proportion were removed.

Following filtration, counts were loaded into a Seurat object using Seurat v4.0.1[42]. Initial clustering was performed using Seurat's reciprocal PCA method of batch correction with *SCTransform* as per authors guidelines

(https://satijalab.org/seurat/articles/integration_rpca.html)[42]. The processed Seurat object was first split by sample and data integration was performed on the first 50 principal components. The principal components of the integrated data were then calculated, which was used to calculate the UMAP embedding for all cells, based on 50 principal components.

To identify clusters, a nearest neighbour graph was constructed on the integrated data, using 50 principal components. Clusters were then identified via the *FindClusters* function using a resolution of 0.6. For the Smillie et al.[22], data analysis, we utilised a k value of 250, as this is concordant with the authors analysis and provided the greatest confidence in identification of clusters than other k values. Concordance was tested by enrichment against the authors cluster markers. For the Elmentaite et al.[23], analysis, we utilised a k value of 20 as any value greater than this resulted in failure to detect the tuft-cell resembling cluster.

To identify potential doublets in the filtered dataset, we utilised *DoubletFinder* v2.0.3[43] as per the authors guidelines (https://github.com/chris-mcginnis-ucsf/DoubletFinder). The fully filtered dataset was then re-used as input for integration, dimensionality reduction and clustering as described above.

To test the robustness of the clusters we identified, we performed pairwise differential gene expression analysis by a receptor operator curve test using Seurat's *FindMarkers* function. In order to only merge clusters which were extremely similar, without over-clustering, we defined similar clusters as those with less than 30 differentially expressed genes with area under the curve score of 0.6 differentiating them. We found no clusters that fall below this threshold and so did not modify our initial clustering in either dataset.

Pan-cluster WGCNA. To analyse correlation of gene expression across all cells from the filtered Smillie et al.[22], data, we first subset for the nominally significant ($p < 0.01$) 11q23.1 trans-eQTLs previously reported[11]. The relative pseudo-bulked expression was then calculated by: (i) summing reads for each gene across all cells within samples, (ii) recombining the summed reads into a non-normalised bulk matrix, (iii) log normalising using TMM normalised size factors, calculated using *edgeR* v3.32.1[44]. Log-TMM normalised bulk expression was then z-scored across genes before analysis.

To perform the network analysis, we used *WGCNA* v1.69[20]. First, *POU2AF2*, *COLCA1* and *COLCA2* pseudo-bulk expression was extracted. A soft threshold of 14 was then selected after computation of mean connectivity and scale free topology, using the recommended 'powerEstimate'. A signed adjacency matrix was then calculated, which was subsequently used to calculate a topological overlap matrix (TOM). Modules were defined by using a dynamic tree cut of hierarchically clustered gene expression by average distances. We did not find any modules with a height of module separation below 0.25 and so did not merge any module. Module eigengenes were then calculated and their correlation with binarized sex, batch, site, and relative, pseudo-bulked expression of *POU2AF2, COLCA1* and *COLCA2* was then assessed. Correlation p-values were multiple testing corrected by the Benjamini–Hochberg method[45]. To visualise the blue module hub genes, a network object was generated from the TOM of blue module genes using *network* v1.17.1[46]. Non-connected genes, along with adjacencies < 0.3, were then removed and remaining genes were plotted using *ggplot2* v3.3.5[47].

Gene set enrichment analysis. All gene set enrichment analysis was carried out using R package fgsea v1.14.0[48]. Genes were ranked by their module membership, gene significance for *POU2AF2* expression, or log fold change of differential expression as stated. In the case where multiple gene sets were being tested, i.e. cluster 11 markers against all Smillie et al.[22], putative markers, p-values were multiple testing corrected by the false discovery rate method.

Calculation of cluster markers. To calculate the markers of each of our own and the Smillie et al.[22], clusters, we first generated a log-transcripts per 10,000 expression matrix for the expression of genes within each cell. This was done so that markers could be calculated using expression values that were not affected by the relative expression of genes within this dataset and so were more applicable to future use. Markers were identified by differential expression of genes within each cluster and all other clusters combined, using MAST v1.160[49].

Analysing the variability of trans-eQTL targets within clusters. To analyse the variability of the expression of 11q23.1 trans-eQTL targets within each cluster, we applied the pseudo-bulking approach described in *Pseudo-bulk WGCNA* to each cluster independently. To make expression variability comparable across samples and clusters, the expression was z-scored across samples.

To analyse expression variability at the single level, we utilised Seurat's *FindVariableFeatures* function and variance stabilising transformation. In concordance with the identification of many trans-eQTL targets identified as markers, their mean expression was very low in several clusters within each dataset. For within dataset comparison of variance, we therefore utilised the raw variance, not normalising for mean expression. For the between dataset comparison of variance, we utilised the normalised variance values. The probability of 11q23.1 eQTL target variation across datasets was calculated by 100,000 permutations of normalised variance values.

Genotyping of Elmentaite et al.[23] samples. To identify the rs3087967 genotype of Elmentaite et al.[23], samples, we utilised two variant calling approaches. These were chosen based on findings from a recent review which identified these methods as the most sensitive for this purpose[50]. Freebayes[24] was used with default settings, genotyping over a 10 bp region including rs3087967. Bcftools[25] variant calling was performed on chromosome 11 using a minimum base quality of 30, disabling read-pair overlap detection and not discarding anomalous pairs.

Sample group definition. In the absence of genotype data for Smillie et al.[22], patients, we grouped samples into high and low expressors of a *POU2AF2*-associated signature by defining the hub genes of the pseudo-bulk WGCNA blue module. These were defined by a module membership (MM) > 0.7, intramodular connectivity (kIM) > 0.7 and network adjacency > 0.3. Samples were then hierarchically clustered by their relative pseudo-bulked expression of these genes using complete distance. We tested the robustness of our sample grouping by bootstrapping, selecting 10 random genes 10,000 times, and counting the number of times this exact separation was achieved—which was 550 times.

Intra-cluster WGCNA. To agnostically identify gene–gene relatedness within cluster 11, we performed WGCNA as described (see *Pseudo-bulk WGCNA*), selecting only the 5000 most variable genes using Seurat's *FindVariableFeatures* and variance stabilising transformation. The scale free topology threshold utilised was 6, as per the 'power estimate'. p-values were corrected for multiple testing as before.

Module preservation analysis. To analyse the preservation of the relatedness of *POU2AF2*-correlated 11q23.1 trans-eQTL targets, we defined an auxiliary module, comprised of the 12 trans-eQTL targets correlated with *POU2AF2* expression (cor > 0.5, $p < 0.05$) within the 'cluster 11 black' module. Pseudo-bulked expression of each cluster was calculated and subset for the 5000 most variable genes as before, see *Pseudo-bulk WGCNA* and *Intra-cluster WGCNA*. The expression from each cluster was then raised to the same scale free topology threshold as cluster 11. The preservation of all modules, including the auxiliary module, was calculated in a pairwise manner with cluster 11, following the author's tutorial (https://horvath.genetics.ucla.edu/html/CoexpressionNetwork/ModulePreservation/Tutorials/)[20]. To summarise module preservation results, we extracted cor.kME values—the correlation of auxiliary module genes with the auxiliary module eigengene, and equivalent results from cluster 11. We also extracted the p-value for this correlation (log.p.cor.kME), which was subsequently un-logged. The mean connectivity of auxiliary module genes with the module eigengene was also calculated within each cluster. Pairwise correlations ($p < 0.05$) between the expression of the auxiliary module genes with one another and *POU2AF2* expression were plotted using corrplot[51].

Linear modelling of cluster abundance and pseudo-bulked gene expression. Univariate linear modelling of cluster 11 abundance was performed on the TMM normalised, non-z-scored pseudo-bulked expression matrix, so that results are more relevant to further study. We fitted a linear model to all genes and performed empirical Bayes moderation using *limma* v3.46.0[52]. p-values were adjusted by Benjamini–Hochberg multiple testing correction[45].

Differential abundance testing. Differential abundance testing was performed using miloR v0.99.8[26]. To mitigate any potential for package-specific artefacts of the analysis, we first re-generated the k nearest neighbour graph using 250 nearest neighbours from the integrated expression. The graph was then built using 4 PCA components. Differential abundance testing was carried out using a Quasi-Likelihood F-test on TMM normalised cell proportions. Differentially abundant neighbourhood results were then annotated for their majority cluster proportion and those comprised of fewer than 80% of the majority cluster were removed.

Human subject inclusion. All the data utilised in this study has been previously published. Informed consent and approval for human subjects from Smillie et al.[14], was obtained from the Prospective Registry in Inflammatory Bowel Disease Study at Massachusetts General Hospital (PRISM:2004P001067). For human subjects in Elmentaite et al.[23], informed consent was obtained from all human participants (reference 15/EE/0152, East of England—Cambridge South Research Ethics Committee). All methods were performed in accordance with the relevant guidelines and regulations.

Data availability

Fastq files are not publicly available for the Smillie et al.[22], data. Raw gene-level expression counts of healthy colonic epithelial cells, published by Smillie et al.[22], were obtained from Single Cell Portal using accession code SCP259. Fastq files for the second dataset, published by Elmentaite et al.[23], were obtained from https://www.ebi.ac.uk/arrayexpress using accession code E-MTAB-9543. All code used for the analysis performed in this study is available at https://github.com/BradleyH017/Harris_et_al_WGCNAscRNA.

Received: 3 June 2022; Accepted: 2 August 2022
Published online: 10 August 2022

References

1. Rawla, P., Sunkara, T. & Barsouk, A. Epidemiology of colorectal cancer: Incidence, mortality, survival, and risk factors. *Gastroenterol. Rev.* **14**, 89 (2019).
2. Cancer Research UK. *Bowel Cancer Statistics*. https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/bowel-cancer#heading-Zero (2018). Accessed Feb 2022.
3. Graff, R. E. *et al.* Familial risk and heritability of colorectal cancer in the nordic twin study of cancer. *Clin. Gastroenterol. Hepatol.* **15**, 1256–1264 (2017).
4. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: Current insights and future perspectives. *Nat. Rev. Cancer* **17**, 692 (2017).
5. Jasperson, K. W., Tuohy, T. M., Neklason, D. W. & Burt, R. W. Hereditary and familial colon cancer. *Gastroenterology* **138**, 2044 (2010).
6. Huyghe, J. R. *et al.* Discovery of common and rare genetic risk variants for colorectal cancer. *Nat. Genet.* **51**, 76 (2019).

7. Law, P. J. *et al.* Association analyses identify 31 new risk loci for colorectal cancer susceptibility. *Nat. Commun.* **10**, 2154 (2019).
8. Hulur, I. *et al.* Enrichment of inflammatory bowel disease and colorectal cancer risk variants in colon expression quantitative trait loci. *BMC Genomics* **16**, 138 (2015).
9. Closa, A. *et al.* Identification of candidate susceptibility genes for colorectal cancer through eQTL analysis. *Carcinogenesis* **35**, 2039 (2014).
10. Loo, L. W. M., Lemire, M. & le Marchand, L. In silico pathway analysis and tissue specific cis-eQTL for colorectal cancer GWAS risk variants. *BMC Genomics* **18**, 381 (2017).
11. Vaughan-Shaw, P. G. *et al.* Differential genetic influences over colorectal cancer risk and gene expression in large bowel mucosa. *Int. J. Cancer* **149**, 1100 (2021).
12. Tenesa, A. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21. *Nat. Genet.* **40**, 631 (2008).
13. Biancolella, M. *et al.* Identification and characterization of functional risk variants for colorectal cancer mapping to chromosome 11q23.1. *Hum. Mol. Genet.* **23**, 2198 (2014).
14. Smillie, C. *Functional Characterisation of the 11q23.1 Colorectal Cancer Risk Locus* (The University of Edinburgh, 2015).
15. Haber, A. L. *et al.* A single-cell survey of the small intestinal epithelium. *Nature* **551**, 333 (2017).
16. Kaske, S. *et al.* TRPM5, a taste-signaling transient receptor potential ion-channel, is a ubiquitous signaling component in chemosensory cells. *BMC Neurosci.* **8**, 49 (2007).
17. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217 (2007).
18. Michaelson, J. J., Loguercio, S. & Beyer, A. Detection and interpretation of expression quantitative trait loci (eQTL). *Methods* **48**, 265 (2009).
19. Richards, A. L. *et al.* Schizophrenia susceptibility alleles are enriched for alleles that affect gene expression in adult human brain. *Mol. Psychiatry* **17**, 193 (2012).
20. Langfelder, P. & Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC Bioinform.* **9**, 559 (2008).
21. Blackmur, J. P. *et al.* Gene co-expression network analysis identifies vitamin D-associated gene modules in adult normal rectal epithelium following supplementation. *Front. Genet.* **12**, 78397 (2022).
22. Smillie, C. S. *et al.* Intra- and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell* **178**, 714 (2019).
23. Elmentaite, R. *et al.* Cells of the human intestinal tract mapped across space and time. *Nature* **597**, 250 (2021).
24. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing (2012).
25. Danecek, P. *et al.* Twelve years of SAMtools and BCFtools. *Gigascience* **10**, 008 (2021).
26. Dann, E., Henderson, N. C., Teichmann, S. A., Morgan, M. D. & Marioni, J. C. Differential abundance testing on single-cell data using k-nearest neighbor graphs. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-021-01033-z (2021).
27. Luo, Y. *et al.* Single-cell transcriptome analyses reveal signals to activate dormant neural stem cells. *Cell* **161**, 1175 (2015).
28. Wu, C.-L. *et al.* Single cell transcriptomic analysis of human pluripotent stem cell chondrogenesis. *Nat. Commun.* **12**, 362 (2021).
29. Gerbe, F. *et al.* Distinct ATOH1 and Neurog3 requirements define tuft cells as a new secretory cell type in the intestinal epithelium. *J. Cell Biol.* **192**, 767 (2011).
30. Yamaga, Y. *et al.* Gene expression profile of Dclk1+ cells in intestinal tumors. *Digest. Liver Dis.* **50**, 1353 (2018).
31. Szczepanski, A. P., Tsuboyama, N., Zhao, Z. & Wang, L. POU2AF2/POU2AF2 functions as a co-activator of POU2F3 by maintaining chromatin accessibility and enhancer activity. *BioRxiv.* https://doi.org/10.1101/2022.03.17.484753 (2022).
32. Wu, X. S. *et al.* OCA-T1 and OCA-T2 are coactivators of POU2F3 in the tuft cell lineage. *Nature*. https://doi.org/10.1038/s41586-022-04842-7 (2022).
33. Gerbe, F. *et al.* Intestinal epithelial tuft cells initiate type 2 mucosal immunity to helminth parasites. *Nature* **529**, 226 (2016).
34. Middelhoff, M. *et al.* Dclk1-expressing tuft cells: Critical modulators of the intestinal niche? *Am. J. Physiol. Gastrointest. Liver Physiol.* **313**, 285 (2017).
35. Banerjee, A. *et al.* Succinate produced by intestinal microbes promotes specification of tuft cells to suppress ileal inflammation. *Gastroenterology* **159**, 2101 (2020).
36. Yi, J. *et al.* Dclk1 in tuft cells promotes inflammation-driven epithelial restitution and mitigates chronic colitis. *Cell Death Differ.* **26**, 1656 (2019).
37. Hoffman, M. T. *et al.* The gustatory sensory G-protein GNAT3 suppresses pancreatic cancer progression in mice. *Cell. Mol. Gastroenterol. Hepatol.* **11**, 349 (2021).
38. DelGiorno, K. E. *et al.* Tuft cells inhibit pancreatic tumorigenesis in mice by producing prostaglandin D2. *Gastroenterology* **159**, 1866 (2020).
39. Kjærgaard, S. *et al.* Decreased number of colonic tuft cells in quiescent ulcerative colitis patients. *Eur. J. Gastroenterol. Hepatol.* **33**, 817 (2021).
40. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
41. Griffiths, J. A., Richard, A. C., Bach, K., Lun, A. T. L. & Marioni, J. C. Detection and removal of barcode swapping in single-cell RNA-seq data. *Nat. Commun.* **9**, 2667 (2018).
42. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573 (2021).
43. McGinnis, C. S., Murrow, L. M. & Gartner, Z. J. DoubletFinder: Doublet detection in single-cell RNA sequencing data using artificial nearest neighbors. *Cell Syst.* **8**, 329 (2019).
44. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288 (2012).
45. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc.* **57**, 289–300 (1995).
46. Butts, C. T. Network: A package for managing relational data in R. *J. Stat. Softw.* https://doi.org/10.18637/jss.v024.i02 (2008).
47. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).
48. Korotkevich, G. *et al.* Fast gene set enrichment analysis. *BioRxiv.* https://doi.org/10.1101/060012 (2021).
49. Finak, G. *et al.* MAST: A flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
50. Liu, F. *et al.* Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biol.* **20**, 242 (2019).
51. Wei, T. & Simko, V. *"Corrplot": Visualization of a Correlation Matrix (Version 0.84)* (2017).
52. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).

Author contributions

Conceptualisation: B.T.H., V.R., M.G.D., S.M.F. Methodology: B.T.H., V.R., S.M.F., M.G.D., A.O., J.P.B., K.D., M.T., P.G.V.S. Investigation: B.T.H. Original draft writing: B.T.H. Review & Editing: B.T.H., V.R., J.P.B., K.D., M.T., A.O., M.G.D., S.M.F. Funding Acquisition: S.M.F., M.G.D., F.V.N.D. Supervision: V.R., M.G.D., S.M.F. All authors read and approved the final manuscript.

Funding
This work was supported by Cancer Research UK (CRUK) PhD studentship at the Edinburgh CRUK Cancer Research Centre (Bradley T. Harris/Susan M. Farrington) and CRUK programme Grant DRCPGM\100012 (Malcolm G. Dunlop/Susan M. Farrington). James P. Blackmur was supported by an ECAT-linked CRUK ECRC Clinical training award (C157/A23218). Peter Vaughan-Shaw was supported by a NES SCREDS clinical lectureship, MRC Clinical Research Training Fellowship (MR/M004007/1), a Research Fellowship from the Harold Bridges bequest and by the Melville Trust for the Care and Cure of Cancer.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-17887-5.

Correspondence and requests for materials should be addressed to S.M.F.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.