Epigenetics Communications

RESEARCH

Open Access

Leukocyte-specific DNA methylation biomarkers and their implication for pathological epigenetic analysis

M. J. Dunnet[1], O. J. Ortega-Recalde[1], S. A. Waters[2,3,4], R. J. Weeks[5], I. M. Morison[5] and T. A. Hore[1*]

Abstract

Background: Distinct cell types can be identified by their DNA methylation patterns. Much research over the last decade has focused on DNA methylation changes in cancer or the use of cell-free circulating DNA in plasma to identify damaged tissue in cases of trauma or organ transplantation. However, there has been little research into the differential methylation patterns between leukocytes and other tissues and how they can be used as a detection tool for immune activity in a range of contexts.

Results: We have identified several loci that are fully methylated in leukocytes but virtually devoid of methylation in a range of other mesoderm-, ectoderm-, and endoderm-derived tissues. We validated these biomarkers using amplicon-bisulphite-sequencing on saliva and in vitro mixing of peripheral blood mononuclear cells and intestinal organoid cells combined at a defined range of ratios. Interestingly, these methylation biomarkers have previously been identified as altered in various inflammatory diseases, including Alzheimer disease, inflammatory bowel disease, and psoriasis. We hypothesise this is due to leukocyte infiltration rather than being a feature of the diseased cells themselves. Moreover, we show a positive linear relationship between infiltrating leukocytes and DNA methylation levels at the HOXA3 locus in six cancer types, indicative of further immune cell infiltration.

Conclusions: Our data emphasise the importance of considering cellular composition when undertaking DNA methylation analysis and demonstrate the feasibility of developing new diagnostic tests to detect inflammation and immune cell infiltration.

Keywords: DNA methylation, Biomarker, Leukocytes, Inflammation, Cancer, Alzheimer disease, Saliva, Deconvolution, HOXA3, MAP4K1

Introduction

Many epigenetic processes are used to modulate gene expression; however, DNA methylation is unique because it can transmit biological memory over long periods of time. In vertebrates, the majority of DNA methylation is found on cytosine, specifically at cytosine-guanine (CpG) dinucleotides. CpG is a palindromic sequence, and as such, methyltransferases, such as DNA methyltransferase 1 (DNMT1), can maintain DNA methylation marks after DNA replication by copying methylation from the cytosine on the template strand to the complementary cytosine on the newly synthesised strand [1]. Because all cells in the body have essentially the same DNA sequence, the cell morphology and function are related to a particular combination of genes that are expressed or repressed. DNA methylation helps regulate gene expression (for example, by preventing transcription factors binding promoters and enhancers [2] or by recruitment of heterochromatin-associated proteins with methyl-binding

*Correspondence: tim.hore@otago.ac.nz

[1] Department of Anatomy, University of Otago, Dunedin, New Zealand
Full list of author information is available at the end of the article

domains [3]), and thus, DNA methylation patterns contribute to defining cellular identity. Accordingly, hierarchical clustering of whole-genome methylation analysis shows that closely related cell types cluster together [4], and many differentially methylated regions remain from earlier developmental cell identity decisions [5].

Unique cell-type-specific DNA methylation patterns can also be used as biomarkers to identify unknown cell-types in forensic and diagnostic settings. For example, cell-free DNA (cfDNA) released from apoptotic and necrotic cells into the bloodstream can be collected with non-invasive blood sampling. Assaying cfDNA for cancer-specific DNA methylation patterns can then be used to detect, monitor, and prognose cancer [6, 7]. cfDNA methylation has also been used to track cell death following traumatic injury or organ transplantation [8–10], and methylation analysis in blood samples has been used to quantify leukocyte subpopulations [11, 12], with application to cancer methylome screening [13, 14].

Although these tissue-type deconvolution studies are powerful and useful in their own right, they rely upon array technology (e.g. Infinium HumanMethylation450 BeadChip), whereby thousands of CpG sites are analysed simultaneously from a single sample. In contrast, relatively simple diagnostic tests, such as those designed to detect inflammation and immune cell infiltration, need only to distinguish blood-derived cells such as leukocytes from other tissue types. As such, complex deconvolution from thousands of loci may not be required [4].

In this study, we first aimed to identify genomic regions that can differentiate between leukocytes and all other cell types and validate them with the use of a rapid, high-throughput bisulphite-PCR assay (Fig. 1). Secondly, we aimed to benchmark these loci against previously published datasets to assess how leukocyte-specific methylation patterns are reported in methylation disease studies. We find two regions within the *HOXA3* and *MAP4K1* loci that are highly methylated in blood-derived cells but unmethylated in all other tissues examined. Following validation using saliva and artificially created DNA mixtures, we found these sites are suitable for quantifying the proportion of leukocytes within heterogeneous tissues.

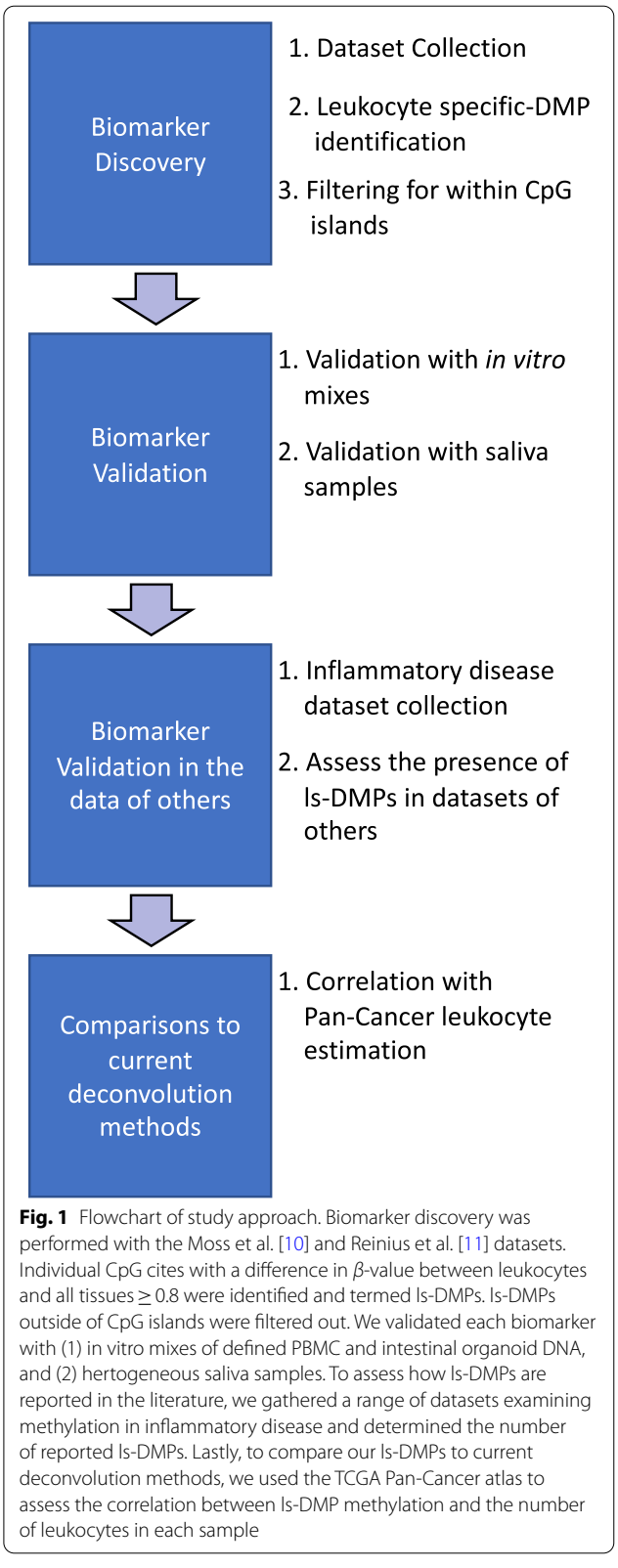

Fig. 1 Flowchart of study approach. Biomarker discovery was performed with the Moss et al. [10] and Reinius et al. [11] datasets. Individual CpG cites with a difference in β-value between leukocytes and all tissues ≥ 0.8 were identified and termed ls-DMPs. ls-DMPs outside of CpG islands were filtered out. We validated each biomarker with (1) in vitro mixes of defined PBMC and intestinal organoid DNA, and (2) hertogeneous saliva samples. To assess how ls-DMPs are reported in the literature, we gathered a range of datasets examining methylation in inflammatory disease and determined the number of reported ls-DMPs. Lastly, to compare our ls-DMPs to current deconvolution methods, we used the TCGA Pan-Cancer atlas to assess the correlation between ls-DMP methylation and the number of leukocytes in each sample

Results

Biomarker discovery

To adequately identify candidate leukocyte-specific DNA methylation biomarkers, we took advantage of the Moss et al. [10] methylation atlas (GEO accession: GSE122126), whereby Illumina Infinium HumanMethylation450 BeadChip array data was created for blood cell preparations, along with a further nine different tissue types that were supposedly free of vasculature and immune populations. For each CpG site in the Moss dataset (423,213 in total),

we calculated a combined mean DNA methylation level (expressed as a β-value between 0 and 1) for all non-leukocyte cells and compared it to the DNA methylation level for leukocytes. The difference between the two means was used to inform the selection of potential biomarkers; CpG sites with a difference of ≥ 0.8 were considered differently methylated (henceforth, we will refer to these as leukocyte-specific differentially methylated positions, or ls-DMPs). In total, 77 ls-DMPs were identified (Table S1). Of these, 19 were highly methylated (methylation > 89%) in leukocytes, while 58 were unmethylated (methylation < 8%). Most of the highly methylated ls-DMPs (15 out of 19) were found in CpG islands (CGIs) and within gene bodies (12 out of 19), while most of the unmethylated sites are in open seas (45 out of 58) (Figure S1a-d). The CGI-located candidate biomarkers were of particular interest because they were likely flanked by additional CpGs providing similar discriminatory power. Subsequently, we only focused on CpG sites located within CGIs (Table 1).

We decided to focus on the only two CGIs that contained multiple ls-DMPs. The first was within the

Table 1 Top ls-DMPs from the Moss dataset located in CpG islands

Probe ID	Leukocyte β-value	Mean non-leukocyte β-value (standard deviation)	Δ β-value	CpG island location	RefSeq Gene ID
cg05258935	0.9626	0.0420 (0.0538)	0.9206	chr19:39086878-39087304	MAP4K1
cg16017089	0.9482	0.0574 (0.0632)	0.8908	chr11:73053648-73054302	ARHGEF17
cg08846870	0.9709	0.0804 (0.0887)	0.8905	chr19:15568027-15569227	RASAL3
cg18856478	0.9237	0.0356 (0.0446)	0.8881	chr1:43814305-43815277	MPL
cg00921266	0.9550	0.0792 (0.1104)	0.8759	chr7:27153187-27153647	HOXA3
cg08101036	0.9638	0.0891 (0.1062)	0.8748	chr7:27153187-27153647	HOXA3
cg14845962	0.9483	0.08120 (0.781)	0.8671	chr12:58119909-58121551	AGAP2
cg10511890	0.0076	0.8655 (0.1793)	0.8579	chr11:47416357-47416598	-
cg17518965	0.0079	0.8583 (0.1431)	0.8503	chr19:3178741-3179986	S1PR4
cg25139229	0.9668	0.1242 (0.1245)	0.8426	chr10:49674243-49674776	ARHGAP22
cg02053964	0.0123	0.8516 (0.1412)	0.8393	chr12:124950705-124950939	NCOR2
cg01681367	0.0352	0.8618 (0.0872)	0.8266	chr16:29675845-29676120	SPN
cg11977716	0.1310	0.9543 (0.0360)	0.8234	chr18:77284291-77285544	NFATC1
cg24006721	0.9415	0.1192 (0.0677)	0.8223	chr1:25255527-25259005	RUNX3
cg02798280	0.9703	0.1556 (0.1480)	0.8147	chr19:39086878-39087304	MAP4K1
cg22987448	0.8896	0.0761 (0.0891)	0.8134	chr19:8591294-8591842	MYO1F
cg05796838	0.9782	0.1699 (0.1658)	0.8083	chr19:17952196-17953474	JAK3
cg08379738	0.9764	0.1701 (0.1380)	0.8064	chr19:6476682-6477127	DENND1C
cg16748008	0.9460	0.1406 (0.1429)	0.8053	chr7:27154999-27155426	HOXA3

List of the top ls-DMP sites identified from the Moss et al. [10] that are within CpG islands. Δ β-value is expressed as the absolute number of the calculated difference. CpG sites within the HOXA3 and MAP4K1 loci have been highlighted

HOXA3 locus (chr7:27,153,187–27,153,647, hg19) and contained two ls-DMPs (cg00921266 and cg08101036). This region also showed strong evolutionary conservation—bisulphite-sequencing in the orthologous mouse region showed a similar methylation pattern (data published by Hon et al. [5]; Figure S2). The second region of interest was within a CGI in exon 26 of *MAP4K1* (chr19:39,086,878–39,087,304, hg19) and also contained two ls-DMPs, cg05258935 and cg02798280. The top CpGs at each locus (cg00921266 and cg05258935) showed similar patterns; very high DNA methylation levels in leukocytes; moderately low methylation levels in adipocytes, lung epithelial cells, and vascular endothelial cells; and virtually no DNA methylation in colonic epithelium, cortical neurons, hepatocytes, and various cells types of the pancreas (Fig. 2a).

To further explore the biological basis for these biomarkers, we examined leukocyte sub-population methylation using previously published array-based datasets [11, 15]. These data show that all adult leukocyte subpopulations are highly methylated at cg05258935 and cg00921266 (Fig. 2b), albeit with a slight reduction in methylation in $CD34^{+}$ stem cells of adult bone marrow (Fig. 2c). In contrast, foetal liver $CD34^{+}$ cells showed significant demethylation compared to adult $CD34^{+}$ (cg00921266, $p = 2.3e^{-5}$; cg05258935, $p = 5.3e^{-9}$) (Fig. 2c). Together, these data suggest that ls-DMPs in *HOXA3* and *MAP4K1* start development in an

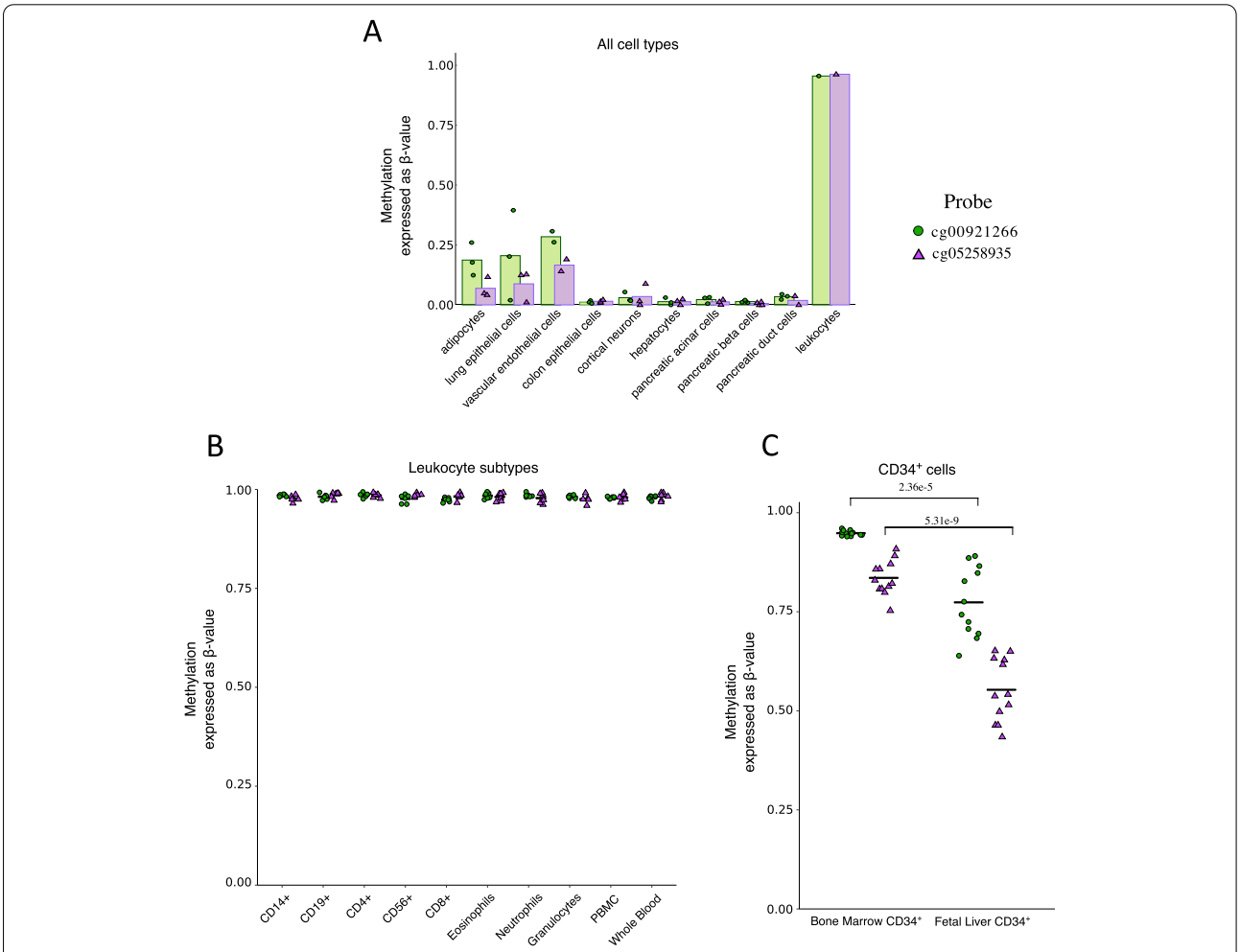

Fig. 2 Marker identification. **a** Bar chart of methylation as a β-value for cg00921266 (HOXA3, green) and cg0528935 (MAP4K1, purple) in the 10 isolated cell populations from the Moss et al. [10] dataset (GEO accession: GSE122126). **b** Dot plot of methylation as a β-value for cg00921266 (HOXA3, green) and cg0528935 (MAP4K1, purple) in 10 different immune cell populations from the Reinius et al. [11] dataset (GEO accession: GSE35069). **c** Dot plot of methylation as a β-value for cg00921266 (HOXA3, green) and cg0528935 (MAP4K1, purple) in CD34.$^{+}$ haematopoietic cells from adult bone marrow and foetal liver

unmodified state, accumulate methylation specifically in the foetal stem cells of the haematopoietic lineage, and maintain this throughout subsequent differentiation, development, and ageing.

Accurate prediction of cell-of-origin identity using DNA methylation

While the datasets identifying these biomarkers are undoubtedly valuable, array technology is limited as a diagnostic platform for several reasons. Firstly, arrays are relatively expensive on a per-sample basis. Second, they rely on single CpG probes to assay methylation at a given site, which risks under-sampling methylation at a given biomarker region. Moreover, array technologies are unable to give information on a single-molecule level. To address this, we decided to pursue a more cost-effective amplicon-bisulphite-sequencing test that takes advantage of the fact that closely related CpGs often share the same DNA methylation state (i.e. methylated or unmethylated). This is because by increasing the number of simultaneously analysed CpG sites, we increase the resolution and discriminatory power for methylation-based biomarkers [9].

To quantify methylation at the *HOXA3* and *MAP4K* loci of interest, we adapted a dual-indexing, four-primer PCR-based assay [16] (see the "Methods" section for an in-depth description). First, we aimed to validate each region of interest and assess the adapted amplicon-bisulphite-sequencing assay's ability to measure DNA methylation levels accurately. Specifically, we wanted to determine if (a) there is amplification bias towards methylated or unmethylated strands and (b) the ability of the assay to discern minority cell types. To accomplish this, we sourced purified DNA from peripheral blood mononuclear cells (PBMCs) as these represent a nucleated cellular portion of blood and should not contain large amounts of serum-derived cfDNA from other tissues. To prepare DNA from cells without any leukocyte or blood-derived origins, we sourced cultured intestinal organoid DNA. Intestinal organoids are grown in precise 3-dimensional culture conditions that support the growth of $Lrg5^{+}$ stem cells and their derivatives [17], therefore avoiding contamination of blood and vasculature. We then mixed the purified DNA from each source in seven precise quantities (Fig. 3a) and subjected each admixture to bisulphite-conversion and the PCR assay.

The pure intestinal organoid DNA had a mean total DNA methylation of 0.83% (standard deviation (sd) = 0.69%) and 1.02% (sd = 0.65%) for *HOXA3* and *MAP4K1*, respectively. In contrast, pure PBMC DNA had a mean total DNA methylation of 94.0% (sd = 0.30%) and 93.1% (sd = 0.21%) for *HOXA3* and *MAP4K1*, respectively. We visualised the DNA methylation of individual reads using heatmaps, where reads run across the row, and each column represents individual CpG methylation status (Fig. 3b, Figure S3a, Figure S4, Figure S5). For *HOXA3*, only a small fraction of reads in the intestinal organoid and PBMC samples were fully methylated (0.37%) or fully demethylated (0.69%), respectively; interestingly, minimal reads contained moderate DNA methylation.

We used a binomial logistic regression model (Figure S6a, Figure S6c) to produce a density plot of the number of methylated cytosine per read for the *HOXA3* and *MAP4K1* loci (Fig. 3c, Figure S3b). The two cell types cluster separately from one another at both loci. We sought to classify the reads from the mixed samples using their DNA methylation pattern. To do this, we constructed a receiver operating characteristic (ROC) curve and determined the optimal classification threshold using the 'pROC' package in R (Figure S6b and S6d). This model determined that *HOXA3* amplicons should be classified as intestinal organoid derived if ≤ 3 of eleven CpGs are methylated or if ≥ 4 are methylated as PBMC derived (TPR = 0.993, FPR = 0.004). For the *MAP4K1* amplicons, if ≤ 2 of eight CpGs were methylated the read was classified as intestinal organoid derived, while ≥ 3 as PBMC derived (TPR = 0.999, FPR = 6.37e − 05). However, because the level of DNA methylation was effectively opposite for each cell type and minimal reads contained moderate levels of DNA methylation, we decided to use a more stringent classification system for each amplicon. Here, a DNA fragment from the *HOXA3* amplicon was classified as intestinal organoid derived if it had ≤ 3 methylated CpGs on it, or PBMC if it was ≥ 6 methylated CpGs, with any reads in between remaining unclassified. Likewise, for the *MAP4K1* amplicon, a read would be classified as intestinal organoid derived if ≤ 2 CpGs were methylated or of PBMC origin if ≥ 6 CpGs were methylated, with any reads in between remaining unclassified.

We applied the classification system to each mixed sample, and both loci showed a remarkable correlation to the amount of input DNA from each cell type (*HOXA3*: $R^2 = 0.999$, *MAP4K1*: $R^2 = 0.9985$) (Fig. 3c, d, Figure S3c-d). Our results also suggest that the specific bisulphite PCR primers we used do not have any significant bias towards highly methylated or unmethylated reads and are highly accurate.

Salivary leukocytes can be deconvoluted from buccal epithelial cells

To further validate these biomarkers with a heterogenous, uncultured tissue sample, we sourced saliva from human donors. We chose saliva because collection is non-invasive, and it contains good numbers of

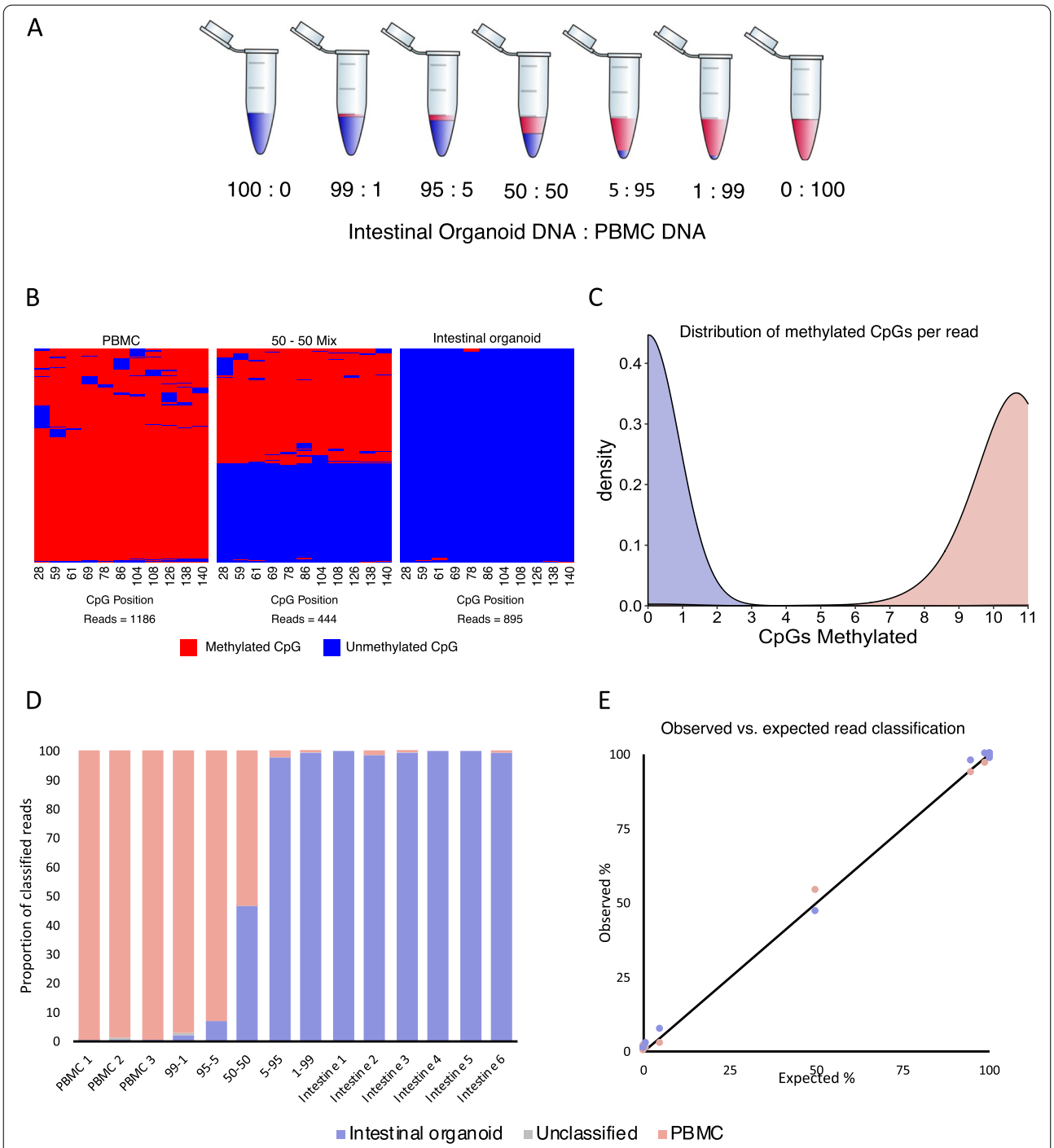

Fig. 3 Deconvolution of in vitro mixed PBMC and intestinal organoid DNA based on HOXA3 DNA methylation. **a** The ratio of Intestinal organoid to PBMC DNA used in the in vitro mixes. **b** Example methylation heatmap outputs of the HOXA3 amplicon for PBMCs (left), 50:50 mix (centre), and intestinal organoid (right). Each row represents an individual read, and each column is a CpG site within the amplicon; column names refer to the CpG position in the read. **c** Density plot for the number of methylated CpG sites per read in pure intestinal organoid (blue) and pure PBMC (red) samples. The *y*-axis represents the probability per unit on the *x*-axis such that the area under the curve for a specific interval is equal to the probability of the number of methylated CpGs in that interval. Bandwidth is 0.45. **d** Stacked bar chart of the proportion of classified for each sample type. Reads were classified as intestinal organoid (blue), PBMC (red), or unclassified (grey). **e** Scatter plot of observed read classification vs expected read classification for intestinal organoid and PBMC

leukocytes secreted from the oral gingiva in addition to non-leukocyte buccal cells sloughed from the cheek epithelium [18]. Buccal cells are large and flat compared to oral leukocytes and are thus easy to identify using standard histological techniques [18]. Furthermore, determining the proportion of salivary leukocytes and buccal epithelium with DNA methylation may be valuable to researchers using saliva samples for epigenetic-based epidemiology studies [19].

To confirm that DNA extracted from leukocytes is methylated and buccal cells unmethylated, we used cellular filtration and flow cytometry to purify respective cell populations from saliva samples. Histological examination of purified cell populations revealed a predicted mean purity of 97.1% for buccal cells and 99.4% for leukocytes (Fig. 4a, b). After performing the dual index bisulphite PCR assay at both the *HOXA3* and *MAP4K1* loci, the vast majority of reads had the expected methylation pattern; however, there was a small sub-population of reads with a largely unmethylated profile in the sorted leukocyte population (2.7% of reads in *HOXA3*; 2.8% in *MAP4K1*) and vice versa for the buccal sorted population (2.3% of reads in *HOXA3*; 2.6% in *MAP4K1*) (Fig. 4d, Figure S7a, Figure S8, Figure S9). We suspect these discrepancies are due to the imperfect isolation of cells as observed in the manual cell counts. We performed the same binomial regression analysis using the isolated buccal and leukocyte cell populations to produce methylation density plots (Fig. 4e, Figure S7b, Figure S10). In this context, the classification cut-off for the two cell types differed from PBMC and intestinal organoid cells. For the *HOXA3* amplicon the cut-off was set as buccal epithelium if five or less CpG sites were methylated or as salivary leukocyte if six or more CpG sites were methylated (TPR = 0.969, FPR = 0.023). For the *MAP4K1* amplicon, reads were classified as buccal epithelium if three or less CpGs were methylated or as a salivary leukocyte if four or more were methylated (TPR = 0.969, FPR = 0.025).

Having confirmed the relationship between methylation of *HOXA3* and *MAPK4* in oral leukocytes and buccal cells, we then sought to quantify their proportion in a mixed saliva sample. The relative proportion of cells in a heterogenous saliva sample was initially assessed using histological examination, and out of a sample of 100 cells, 38% were salivary leukocytes and 62% buccal epithelium (Fig. 4a). Following bisulphite conversion and PCR assay (Fig. 4f, Figure S7) of DNA purified from this same sample, we found that the predicted number of leukocytes was 38.1% (sd = 2.5) and 39.3% (sd = 1.5) for *HOXA3* and *MAP4K1*, respectively. A linear regression model suggests a high degree of accuracy ($R^2_{HOXA3} = 0.998$, $R^2_{MAP4K1} = 0.997$) (Fig. 4g, Figure S7c). These findings indicate that *HOXA3* and *MAP4K1* DNA methylation patterns can accurately deconvolute leukocytes and buccal epithelial cells from a mixed saliva sample.

Independent validation of ls-DMPs in inflammatory disease

Immune cells drive inflammation, and as such, if the ls-DMPs we identified were valid, we would expect them to be overrepresented in methylation datasets featuring inflammatory disease tissue, compared to controls. Using the online database 'EWAS Atlas' (https://ngdc.cncb.ac.cn/ewas/atlas) [20], we discovered 28 associated traits from 35 publications (Table S2) that were associated with our ls-DMPs. The top three traits with the highest number overlapping CpG sites were psoriasis (33 DMPs identified by Chandra et al. [21]), inflammatory bowel disease (IBD) (12 DMPs identified by Agliata et al. [22]), and Alzheimer disease (AD) (12 DMPs identified over three publications [23–25]).

While the EWAS Atlas is an excellent resource, it is limited by the small number of publications in the database (910 at the time of analysis). For that reason, we searched the PubMed database for studies that examined differences in DNA methylation in cases vs controls of the aforementioned traits (Table 2 and Table S3). In total, we found three publications on psoriasis, two on IBD, and seven on AD. Of the 11 publications utilising the Illumina Infinium HumanMethylation450 BeadChip array platform, we found 7 featured at least one differentially methylated region overlapping our ls-DMPs—cg00921266 and cg08101036 from the *HOXA3* locus were present in nine and four datasets, respectively, while cg05258935 and cg02798280 from the *MAP4K1* locus were present in three and four datasets, respectively. Additionally, the *HOXA3* and *MAP4K1* loci were represented in each disease trait at least once (Table S4 and Table S5).

The *HOXA3* locus was vastly over-represented in publications focusing on AD [23, 29, 30, 32, 34], with consistent reports of hypermethylation correlating with AD severity. For example, both De Jager et al. [32] and Smith et al. [30] independently report a 48-kb region spanning the HOXA cluster from *HOXA2* to *HOXA6* as hypermethylated with respect to Braak stage. We hypothesised that this signature resulted from infiltrating leukocytes to the AD brain, a phenomenon known to occur in the ageing human brain and mouse models of AD as the blood–brain barrier becomes leaky [35, 36]. We replicated the DMP discovery pipeline used by Smith et al. [30] for the HOXA cluster (bottom panel, Fig. 5b) and compared this to leukocyte and cortical neuron methylation in the Moss dataset (GEO

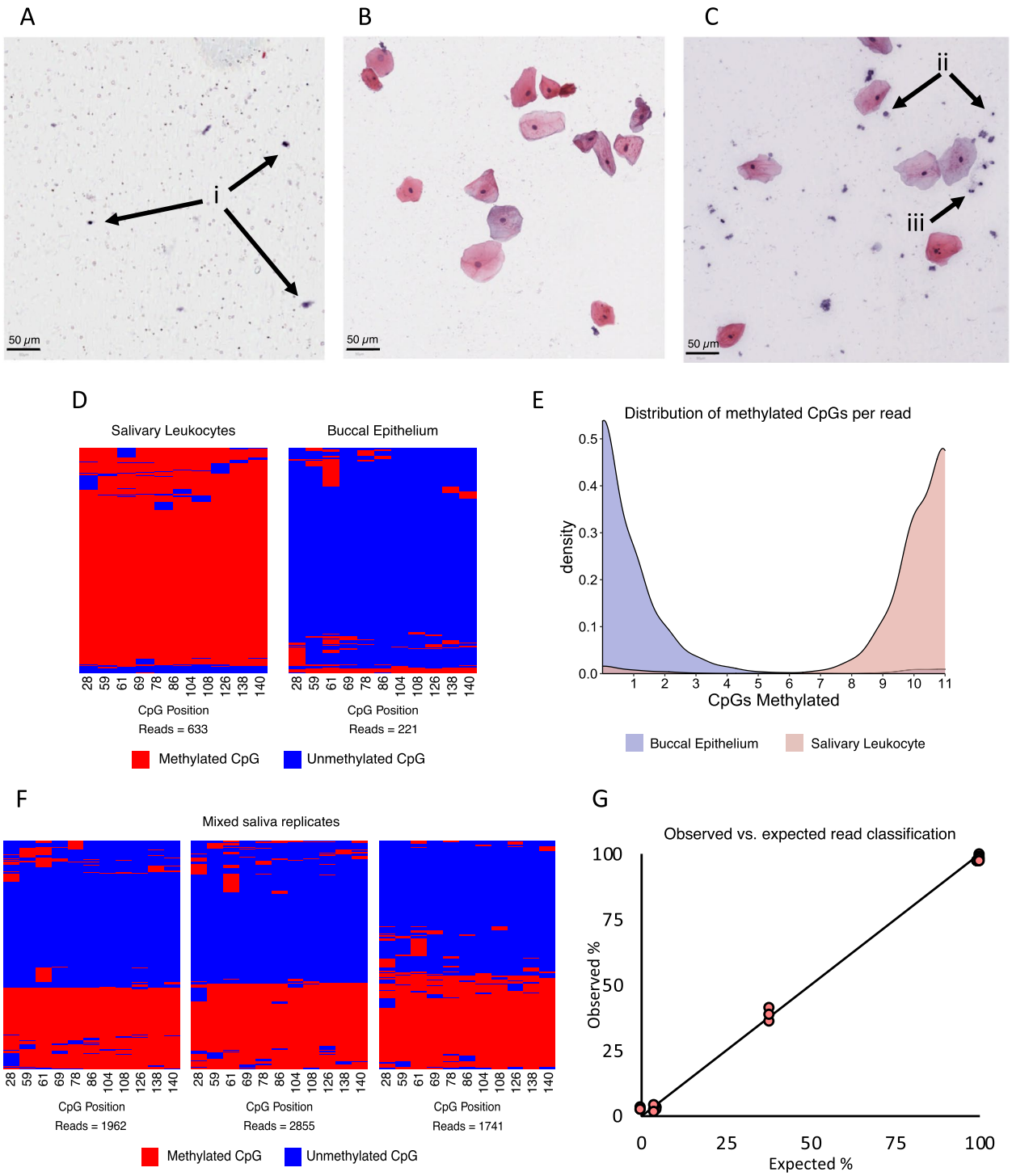

Fig. 4 Deconvolution of the cellular components of saliva. **a**–**c** light microscopy images of H&E stained **a** salivary leukocytes (salivary leukocytes are labelled i), **b** buccal epithelial cells, and **c** mixed saliva (mononuclear cells are labelled ii, and polymorphonuclear cells are labelled iii). **d** Example of methylation heatmaps for isolated salivary leukocytes and buccal epithelium. Each row represents an individual read, and each column is a CpG site within the amplicon; column names refer to the CpG position in the read. **e** Density plot (as in Fig. 3c) for the number of methylated CpG sites per read in isolated buccal epithelium (blue) and salivary leukocyte (red) samples (bandwidth = 0.45). **f** Methylation heatmaps (as explained in Fig. 3b) for mixed saliva samples performed in triplicate. **g** Scatter plot of observed read classification from methylation analysis vs expected read classification from microscopic measurement of the proportion of salivary leukocytes

Table 2 List of publications used in the analysis of ls-DMPs and inflammatory disease

Publication	Trait	Tissue	Data type	HOXA3 or MAP4K1 included
Chandra et al., 2018 [21]	Psoriasis	Skin	450 k methylation bead array	HOXA3
Zhou et al., 2015 [26]	Psoriasis	Skin	450 k methylation bead array	HOXA3 and MAP4K1
Verma et al., 2018 [27]	Psoriasis	Skin	Reduced representation bisulfite sequencing	HOXA3
Agliata et al., 2020 [22]	IBD	Colonic mucosa and purified IEC	450 k methylation bead array meta-analysis	HOXA3 and MAP4K1
Harris et al., 2020 [28]	IBD	Colonic mucosa	450 k methylation bead array	HOXA3
Zhang et al., 2020 [23]	AD	PFC	450 k methylation bead array meta-analysis	HOXA3 and MAP4K1
Gasparoni et al., 2018 [29]	AD	PFC	450 k methylation bead array	HOXA3 and MAP4K1
Smith et al., 2018 [30]	AD	PFC and STG	450 k methylation bead array meta-analysis	HOXA3 and MAP4K1
Li et al., 2020 [31]	AD	STG and IFG	EPIC array	HOXA3
Jager et al., 2014 [32]	AD	PFC	450 k methylation bead array	HOXA3
Altuna et al., 2019 [33]	AD	Hippocampus	450 k methylation bead array and bisulfite sequencing	HOXA3 and MAP4K1
Smith et al., 2021 [34]	AD	PFC, STG, MTG, EC, Cerebellum	450 k methylation bead array meta-analysis	HOXA3 and MAP4K1

List of publications used in the analysis of ls-DMPs and inflammatory disease. *IDB* inflammatory bowel disease, *AD* Alzheimer disease, *IEC* intestinal epithelial cells, *PFC* prefrontal cortex, *STG* superior temporal gyrus, *IFG* inferior frontal gyrus, *MTG* middle temporal gyrus, *EC* entorhinal cortex

accession: GSE122126). Overlaying the plots clearly shows that leukocytes possess a high degree of DNA methylation at the HOXA locus, where cortical neurons do not. This difference in DNA methylation mirrors the exact genomic location of hypermethylation present in AD, suggesting a high degree of correlation between the two observations (Fig. 5b). Indeed, comparing cortical neurons and leukocytes from the Moss dataset shows that five ls-DMPs within the two *HOXA3* CGIs previously described in Table 1 (chr7:27,153,187–27,153,647 and chr7:27,154,999–27,155,426) have β-value differences of ≥ 0.9 (Table S6). Assuming we could extend the principles of our bisulphite-sequencing test to array data, we performed a simple immune cell quantification with Braak stage 0 and six samples from the Smith et al. [30] data using cg00921266. We calculate that as a result of changes in a Braak stage 6 brain, an additional 13.9% (sd = 6.6%) of total cells are leukocyte-derived. (Figure S11).

Interestingly, many of the publications we examined did not even mention leukocyte infiltration as a confounding factor in their quest to find disease biomarkers, despite examining inflammatory diseases [21, 26, 29–31]. Others mentioned leukocyte infiltration as a possibility or discussed cellular heterogeneity [22, 23, 28, 32, 34] but did not control for this variability. In some cases, this occurred in meta-analyses where the authors examined previously collected data and conducting additional experimental controls was not possible [22, 23, 34]. Only one publication investigated the proportion of leukocytes in their samples; however, these data were not shown [27].

In summary, the presence of many ls-DMPs in the methylation datasets of multiple inflammatory diseases further validates their usefulness as pan-leukocyte biomarkers. Our analysis also highlights the minimal use of controls for cellular composition these studies employed. Lastly, we show a strong correlation between the previously unexplained hypermethylation of the HOXA cluster in AD and the DNA methylation differences between leukocytes and brain tissue. We suggest that leukocyte infiltration is driving the change in methylation rather than the alteration in the methylation status of resident cell types.

Comparisons to immune cell deconvolution techniques in cancer

The immune system has a complex and key role in responding to tumours, and in some cases, inflammation can bring about tumourigenesis [13, 37]. A landmark study by Thorsson et al. [13] quantified the immune response to a large array of different cancer types. Thorsson et al.

(See figure on next page.)

Fig. 5 HOXA3 and MAP4K1 in inflammatory disease. **a** Flow chart of online data selection methodology. **b** Methylation expressed as a *β*-value for leukocytes (first panel, red), cortical neurons (second panel, blue), and the difference in *β*-value between the two (third panel, black) for the HOXA cluster. Data from the Moss et al. [10] dataset GEO accession: (GSE122126). The bottom panel is a reconstructed mini-Manhattan plot of the HOXA cluster methylation with respect to Braak stage classification in Alzheimer disease. *P*-values were generated using a linear regression model of DNA methylation with respect to each Braak stage classification (0–6). Red circles indicate an increase in methylation (*β*-value) by ≥ 0.1 as Braak stage increases, green a decrease ≥ 0.1, and black a change < 0.1. Data from Smith et al. [30] (GEO accession: GSE80970)

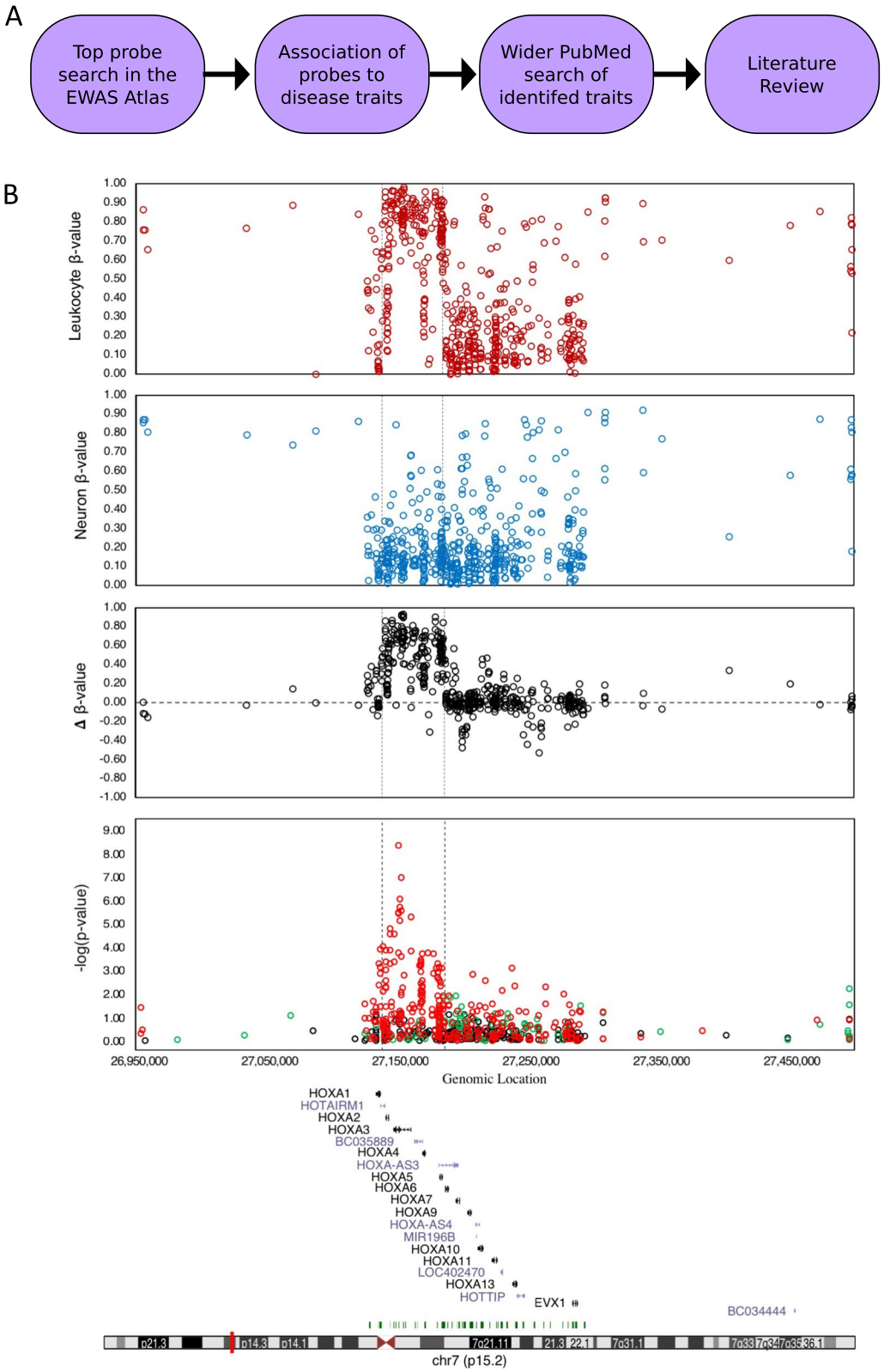

Fig. 5 (See legend on previous page.)

used a relativity complex deconvolution method requiring the Illumina Infinium HumanMethylation450 BeadChip array; we asked if the ls-DMPs identified in our study could predict the proportion of leukocyte populations within non-blood derived tumours without the use of array technology. Using the TCGA database, we compared the DNA methylation of cg00921266 (*HOXA3*) to the leukocyte estimate by Thorsson et al. [13] for over 8000 individual tumour samples from 30 different tumour types (Fig. 6a). A pan-cancer comparison showed an overall poor correlation ($R^2 = 0.286$); nevertheless, many data points cluster in a 1:1 linear relationship between the leukocyte estimate and cg00921266 methylation. The remaining data sit above this line, suggesting that the number of leukocytes in the tumour sets a baseline DNA methylation at this locus, that epimutations are frequent at the *HOXA3* locus in cancer, and that these are virtually all hypermethylation events. Breast invasive carcinoma, colon adenocarcinoma, hepatocellular carcinoma, and cutaneous melanoma are individual examples of this (Figure S12). This is consistent with other reports highlighting hypermethylation at the HOXA cluster genes in cancer [38–41]. While the correlation between cg00921266 and the leukocyte estimate by Thorsson et al. was difficult to see for many individual tumours, six tumour types (uveal melanoma (p-value = 9.70e − 11), thyroid carcinoma (p-value = 2.32e − 08), testicular germ cell tumours (p-value = 5.12e − 08), urothelial bladder carcinoma (p-value = 9.5e − 08), mesothelioma (p-value = 6.28e − 21), and ovarian serous cystadenocarcinoma (p-value = 1.31e − 03)) showed a remarkable 1:1 relationship (Fig. 6b–g). These data suggest that for at least some cancer types, the methylation of a single CpG site (cg00921266) is all that is required for an accurate determination of leukocyte infiltration to the tumour.

Discussion

Using array data, we identified several CpG sites in the *HOXA3* and *MAP4K1* loci as specifically methylated in all major blood cell populations (Fig. 2b). We performed validation experiments using PBMCs and salivary leukocytes; PBMCs are primarily comprised of $CD4^+$ and $CD8^+$ T-cells, and to a lesser extent, monocytes, B-cells, and natural killer cells [42], while 95% of salivary leukocytes are neutrophils [18]. Both of these tissue types showed high methylation in the leukocyte-derived portions. Therefore, we have identified and validated two regions with unique DNA methylation patterns across the majority of mature blood-derived cell types.

The biological significance of HOXA3 and MAP4K1 DNA methylation in blood-derived cells

Although biomarkers do not need to reflect function, correlation with biological function can provide confidence in a marker. HOX genes are a highly conserved family of transcription factors involved in haematopoiesis, particularly of myeloid committed cells [43–45]. The development of definitive haematopoietic cells occurs in the foetal aorta, whereby a transition from endothelium to haematopoietic stem cells occurs [46]. Studies performed in mice show that the medial to late HOXA genes *HOXA5*, *HOXA7*, and *HOXA9* are critical for endothelium to haematopoietic stem cell transition [47], and *HOXA10* is responsible for differentiation towards the myeloid and erythroid lineages [48]. *HOXA3*, unlike the other genes in the HOXA cluster, acts to maintain an endothelial phenotype, and its expression in haematopoietic stem cells can drive them back towards an endothelial lineage [49]. Collectively, these data suggest that while medial and late HOXA genes drive the differentiation of haematopoietic stem cells, the early HOXA gene, *HOXA3*, inhibits this differentiation. The increased DNA methylation observed in the early HOXA genes relative to the medial and late genes (Fig. 5b, top panel) may reflect the silencing of *HOXA3* during haematopoietic stem cell differentiation.

MAP4K1 is a serine/threonine kinase heavily involved in JNK signalling and NF- κB regulation [50]. MAP4K1 appears to dampen the activation of T-cell and B-cell receptors by phosphorylating and inhibiting the activity of important signalling molecules such as SLP-76 or BLNK [50]; additionally, MAP4K1 possesses a dual role in immune cell adhesion, inhibiting the adhesion of lymphocytes while enhancing that of neutrophils [50]. The differentially methylated region within the *MAP4K1* locus is found within the gene body, while the promotor region is consistently unmethylated across cell types. This is consistent with a recent study showing that most intragenic CGIs become methylated in association with transcription [51]. Tissue expression data from the Human Protein Atlas [52] (http://www.proteinatlas.org) of *MAP4K1* shows that it is highly expressed in blood, lymphoid, and gastrointestinal tissues; single-cell expression data show that

(See figure on next page.)

Fig. 6 The relationship between cg00921266 methylation and the proportion of leukocytes in cancers**.** Scatterplots of cg00921266 DNA methylation (expressed as a β-value vs a leukocyte estimate for **a** all cancers combined in the TCGA pan-cancer dataset, **b** thyroid carcinoma, **c** bladder urothelial carcinoma, **d** mesothelioma, **e** testicular germ cell tumours, **f** uveal melanoma, and **g** ovarian serous cystadenocarcinoma. Data was obtained from the TCGA Pan-cancer atlas publication page (https://gdc.cancer.gov/about-data/publications/pancanatlas)

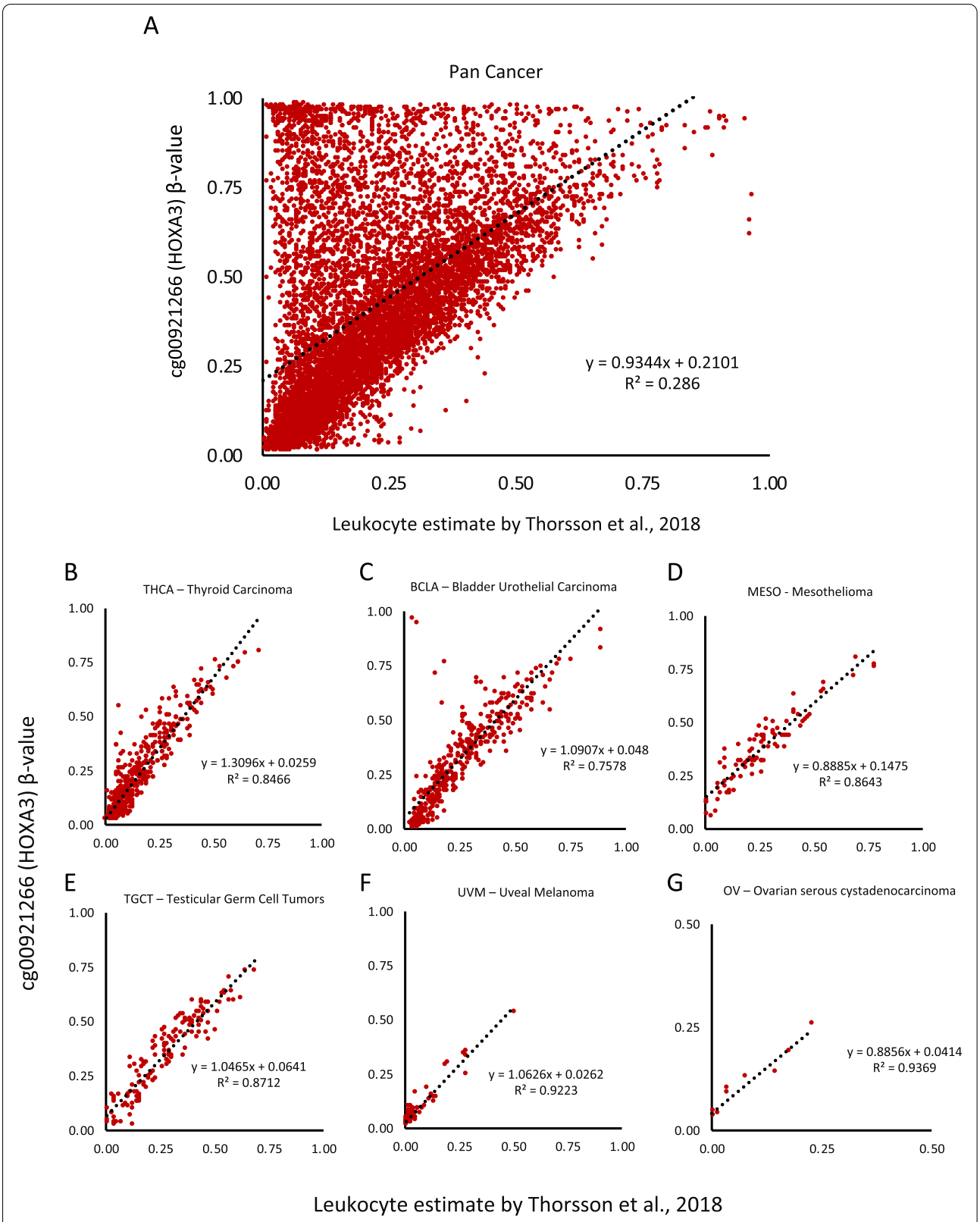

Fig. 6 (See legend on previous page.)

gastrointestinal tract expression is primarily from the result of resident immune cells.

Overall, there is evidence to suggest leukocyte-specific DNA methylation patterns at *HOXA3* and *MAP4K1* have a function within a biological context and is consistent with the finding that methylation is deposited at these two sites early in haematopoietic stem cell development, possibly before bone marrow control of haematopoiesis, and is maintained throughout maturation to adult cells (Fig. 2c).

Deconvolution of the cellular composition in saliva

Saliva is a widely used specimen type for epigenetic epidemiology studies because it is non-invasive, is easily stored for an extended period, and contains ample human DNA [19, 53]. However, saliva-based studies are not without their limitations. A recent review by Langie et al. [53] discusses how the collection method greatly influences the sample's composition. Additionally, the age of individuals also significantly impacts sample heterogeneity [18].

Saliva heterogeneity is a substantial hurdle for epigenetic researchers; DNA methylation differences may reflect the typical biological differences between cell types rather than a disease biomarker. Current cellular deconvolution methods attempt to do this bioinformatically; however, most require reference datasets for comparison. Unfortunately, these methods can only be applied to whole-genome data, and the choice of reference data can significantly affect estimated cell populations [19, 53, 54]. Using our dual-index, 4-primer PCR assay, we have shown that it is possible to precisely determine the proportion of leukocytes to buccal cells in a saliva sample, potentially solving this issue.

Applications of HOXA3 and MAP4K1 DNA methylation in a disease context

We have shown that *HOXA3* and *MAP4K1* can accurately distinguish leukocytes in a mixed-cell system (e.g. intestinal organoids vs PBMCs; salivary leukocytes vs buccal epithelium). Our experiments are a valid proof of concept that high-throughput bisulphite amplicon sequencing has excellent potential for ascertaining the proportion of leukocytes within a tissue sample, although it remains to be seen how it will perform in more complex biopsies or tissue systems. Nevertheless, detecting a large number of ls-DMPs in publications examining a range of inflammatory diseases is encouraging in this regard (Fig. 5).

When we examined the presence of ls-DMPs in the data of others, each study reported at least one DMP in our list of ls-DMPs or reported a DMR that contained one of our ls-DMPs (Table S4 and Table S5). We also show that hypermethylation of the HOXA cluster in Braak stage 6 Alzheimer disease, which is commonly reported [23, 29–34], has a striking overlap with differences in DNA methylation between normal leukocytes and neurons. The coinciding ls-DMPs, and particularly *HOXA3*, suggest that the difference in DNA methylation observed in these studies may result from infiltration of leukocytes rather than a change of DNA methylation in the endogenous cells of affected tissues. We suggest using additional controls in case–control methylation experiments investigating inflammatory diseases to account for the presence of leukocytes. This may include (a) the use of pure cell populations sorted by flow cytometry or equivalent technology, (b) control samples with defined amounts of spiked in leukocyte DNA to simulate leukocyte infiltration, or (c) comparisons to methylation data from healthy tissues and leukocytes to identify cell-type-specific differences in DNA methylation.

We observed many ls-DMPs in methylation datasets examining IBD [22, 28]. Based on array data [10] (Fig. 2a), the amplicon-bisulphite-sequinning data we collected (Fig. 3), and whole genome sequencing data from mice [5] (*HOXA3* only, Figure S2), both the *HOXA3* and *MAP4K1* loci are virtually devoid of DNA methylation in the colonic epithelium. Therefore, it is conceivable that these loci could also be applied to stool samples as intestinal inflammation markers, in particular as a diagnostic tool for IBD. Clinicians categorise IBD as either Crohn's disease or ulcerative colitis; both are characterised by a breakdown of the intestinal mucosal barrier resulting in chronic inflammation [55]. The gold standard for IBD diagnosis is a colonoscopy; however, this is an invasive and time-consuming procedure. Although it remains to be tested, these biomarkers may have utility in this setting and a fully developed diagnostic test may be an alternative to the current calprotectin assay [55, 56].

Using the pan-cancer data from the TCGA database (30 cancers from over 8000 individuals), we have shown that DNA methylation from a single CpG site, cg00921266 (*HOXA3*), has a strong linear correlation with a total leukocyte estimate (previously published by Thorsson et al. [13]) in six different cancer types: uveal melanoma, mesothelioma, testicular germ cell tumours, bladder urothelial carcinoma, ovarian serous cystadenocarcinoma, and thyroid carcinoma. Despite the heterogeneity of cancer cells within a singular tumour, let alone different tumour types [57, 58], our data suggest that differential methylation patterns at the *HOXA3* locus are maintained throughout carcinogenesis in these cancers [59]. As such, an amplicon-based approach test, such as that we have used here, or 450 K Illumina array data from just a single CpG (cg00921266), should provide accurate estimations of leukocyte proportion without complex deconvolution [4].

Conclusions

In conclusion, we have validated two loci within the *HOXA3* and *MAP4K1* regions using a high-throughput amplicon-bisulphite-sequencing approach and accurately measured the proportion of leukocytes within two different contexts. This may be valuable in both a clinical and research setting by removing the need for array-based deconvolution at a lower cost and higher throughput. Importantly, we discovered that many of these ls-DMPs are reported in studies examining the methylation differences in inflammatory diseases, suggesting that the signal is the result of infiltrating leukocytes rather than a change in native cells. Lastly, we show that in a collection of six cancer types, the methylation of a single CpG site (cg00921266) at the *HOXA3* locus correlates highly with a leukocyte estimate performed by Thorsson et al. [13], suggesting a single CpG site or amplicon may be able to determine the proportion of leukocytes in a cancer sample accurately.

Methods

Data acquisition

We employed several Illumina InfiniumHumanMethylation450 BeadChip datasets from the Gene Expression Omnibus and The Cancer Genome Atlas programme (TCGA) for our analysis. All data were downloaded and imported into R-Studio and Microsoft Excel for analysis. We performed initial biomarker discovery with the Moss dataset (GEO accession: GSE122126) [10] and validation of blood-derived methylation patterns with the Reinius dataset (GEO accession: GSE35069) [11]. To examine the methylation of *HOXA3* and *MAP4K1* in different haematopoietic stem cell populations, we utilised the Lessard dataset (GEO accession: GSE56491). Analysis of DNA methylation differences at the HOXA cluster in Alzheimer disease was recreated using dataset produced by Smith et al. (GEO accession: GSE80970) [30]. We obtained the pan-cancer DNA methylation data and corresponding leukocyte estimates from the TCGA Pan-cancer Atlas publication page (https://gdc.cancer.gov/about-data/publications/pancanatlas), respectively [13, 60].

Biomarker discovery

We performed initial biomarker discovery with the dataset published by Moss et al. [10] (GEO accession: GSE122126). This dataset contained methylation data in the form of β-values for 423,213 individual CpG sites across. We excluded all cfDNA and in vitro mix samples, only examining data from sorted, healthy tissue samples. In total, we used 23 of 101 samples in the dataset. We calculated the difference between leukocyte β-values and the mean β-values of all other cells to identify leukocyte-specific methylation patterns. We considered CpG sites with an absolute β-value difference of ≥ 0.8 as ls-DMPs. ls-DMPs outside CpG islands (as described by Illumina Infinium) were discarded to ensure a high density of CpG sites per read. CpG probe locations were determined with the annotation package, 'IlluminaHumanMethylation450kanno.ilmn12.hg19', in R-Studio.

Sample collection and processing

PBMC isolation and DNA purification

PBMCs were isolated using density centrifugation with Ficoll-Paque. Simply, blood was collected into EDTA collection tubes. After transfer to 50-mL Falcon tubes, the blood was mixed with an equal volume of PBS, before careful layering onto Ficoll-Paque solution (room temperature) in a clean, 50 mL Falcon tube. After centrifugation (800 g for 20 min at room temperature, with no braking), the mononuclear cells (on top of Ficoll-Paque layer) were extracted. The PBMCs were then washed several times with RBC lysis buffer (155 mM NH_4Cl, 10 mM $KHCO_3$, 0.1 mM EDTA) to remove contaminating red blood cells. PBMC DNA was then isolated using the QIAAMP DNA Blood mini kit (QIAGEN #QIAG51104).

Intestinal organoid culture and DNA purification

DNA was extracted from cryopreserved intestinal organoids using UltraPure Phenol:Choloroform:Isoamyl Alcohol (25:24:1) (Invitrogen Cat#15,593–031) following the manufacturer's instructions. Isolated DNA was resuspended in RNAse-free water and quantified using Thermo Scientific Nano Drop Spectrophotometer. Intestinal organoids were cultured according to the established protocol [61] from crypts isolated from rectal biopsies as described previously [62]. The participants' provided written informed consent. The sample collection was approved by the Sydney Children's Hospital Ethics Review Board (HREC/16/SCHN/120). Crypts were seeded in Extracellular Matrix (70% matrigel (Growth factor reduced, phenol-free; Corning 356,231) in 24-well plates at a density of ~ 10– 30 crypts in 3×10 µl droplets per well. IntestiCult Organoid Growth Media (STEMCELL Technologies) change was performed every second day and organoids were grown for 7–10 days, harvested from Matrigel and cryopreserved prior to DNA extraction.

Saliva preparation, cell isolation, and DNA purification

We collect 5-mL unstimulated saliva samples as previously described [18]. In the 30 min prior to collection, participants only consumed water. To isolate salivary leukocytes from buccal epithelium, we centrifuged samples at 400 RCF and washed them with 0.01 Molar PBS. Buccal epithelial cells were isolated using fluorescence-activated cell sorting (FACS) using only forward-scatter and side-scatter. Salivary leukocytes were isolated using sequential cellular filtration. Samples were first filtered

through a 40-µm, then a 20-µm mesh filter to exclude buccal cells.

We smeared a fraction of each sample onto microscope slides and stained them with haematoxylin and eosin to assess the cellular composition of both the sorted cell populations and mixed saliva. Cell counts were performed manually, and each slide was viewed under the microscope using a 20 × objective. A field of view was chosen at random where there were appropriate cell numbers, all the cells in the field could be adequately identified (i.e. clumping and overlapping of cells), and there was an acceptable level of staining to identify each cell. We counted a minimum of 100 cells across two fields per slide.

DNA was extracted using the BOMB.bio TNA extraction from mammalian tissue protocol [63]. In short, cells were lysed in 1 × TNES and 0.5 µL of 20 mg mL^{-1} of proteinase K. The lysate was mixed with 6 M GITC, carboxyl-coated magnetic beads in suspended TE, and absolute isopropanol in a ratio of 1:2:2:4, respectively. After a 5-min incubation, samples were applied to a magnet and washed once with absolute isopropanol and twice with 70% ethanol. We eluted the DNA from the magnetic beads with TE buffer and measured the concentration with high sensitivity dsDNA Qubit.

Bisulphite conversion and amplicon sequencing

We performed bisulphite conversion with 200 ng of extracted DNA using the Zymo Research EZ-96 DNA Methylation MagPrep kit per the manufacturer's guidelines and measured DNA concentration with the ssDNA Qubit. We amplified at least 75 ng of converted DNA using the dual-index, four-primer PCR assay (Fig. 7). We performed the reaction over two amplification steps. In the first step, converted DNA was amplified with the KAPA HiFi Uracil + ReadyMix and either 0.1 µM or 0.3 µM of first step primers for *HOXA3* and *MAP4K1* amplicons, respectively. Each reaction was topped up to 25 µL with nuclease-free H_2O. Amplification cycling parameters were 95 °C for 2 min, 23 cycles of 98 °C for 20 s, 59 °C for 10 s, and 72 °C for 20 s. A final elongation step was performed for 5 min at 72 °C. Reactions were placed on ice, and 0.2 µM of the second step primers (Illumina P5 and P7 adapters and TruSeq indexes with added linker sequence, Table S7) was added. Amplification was repeated as above for five additional cycles. We used solid-phase reverse immobilisation of carboxyl-coated magnetic beads suspended in polyethylene glycol to size select for DNA fragments the length of the amplicons [63]. We sequenced the amplicons on the Illumina iSeq100 system.

Primer sequences can be found in Table S7. The HOXA3 amplicon includes cg00921266 and cg08101036. Suitable primers for the MAP4K1 amplicon could not be designed to include cg05258935 or cg02798280; however, the amplicon targets the same CpG island within 250 base pairs of both sites.

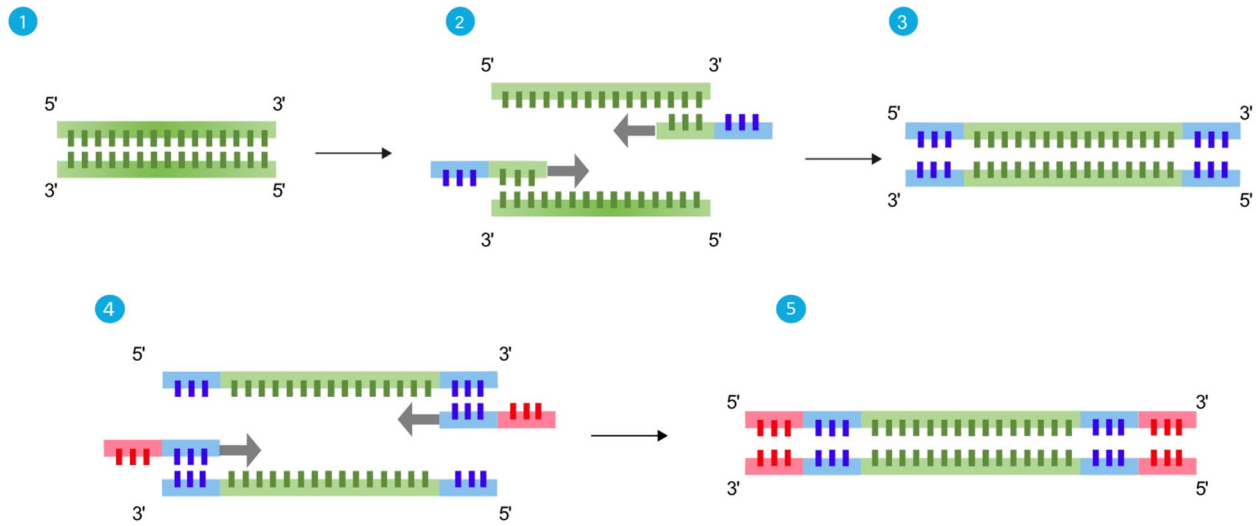

Fig. 7 Diagram of dual index, four-primer amplicon-bisulfite PCR. The bisulfite converted template DNA (1) is amplified in an initial cycle with primers containing a sequence specific region (green) and overhanging 'handle' sequence (blue) (2). After amplification, the handle is incorporated into the amplicon (3). A second amplification cycle is completed with primers complementary to the handle (blue) and an overhang consisting of a multiplex sequencing index and Illumina adapters (red) (4) such that the indexes and adapter are incorporated into the final PCR product (5)

Data processing and statistical analyses

We removed adapter sequences from each read using Cutadapt and TrimGalore [64]. Using Bismark [65], we mapped the amplicon reads to a custom 'genome' containing only the amplicon sequences. The sequences used for mapping were obtained from the UCSC genome browser hg38 (*HOXA3*: chr7:27,113,957–27,114,300, *MAP4K1*: chr19:38,596,411–38,596,696). Whole-genome bisulfite sequencing data of 19 different mouse tissues obtained from Hon et al. [5] were mapped to the MGSCv37 (mm9) mouse reference genome using Bismark and visualised using SeqMonk.

We produced the heatmaps, density plots, cell-type-calling analysis, and linear regression using custom R-scripts in R-Studio. We performed the logistic regression analysis and ROC curve construction in R-Studio using the pROC package.

We estimated the proportion of infiltrating leukocytes in the prefrontal cortex of individuals with Alzheimer disease using the methylation of cg00921266. To do this, we employed the formula: $L_{est} = \frac{m_6 - m_0}{m_L}$, where the L_{est} is the leukocyte estimate, m_6 is the methylation in the prefrontal cortex at Braak stage 6, m_0 the methylation at Braak stage 0, and m_L the methylation in leukocytes based on the Moss dataset.

Selection of online datasets for analysis of inflammatory disease

We searched for all ls-DMPs within the EWAS Atlas depository (https://ngdc.cncb.ac.cn/ewas/atlas) to determine associated disease traits correlating with hypo- or hypermethylation. We summed the number of unique ls-DMPs associated with each disease trait and focused a more comprehensive literature search on the top three diseases using the PubMed database. Our specific search terms were 'DNA methylation' and either 'psoriasis', 'inflammatory bowel disease', or 'Alzheimer disease'. We included publications from 2010 onwards and where analysis of affected tissues was examined (i.e. skin, intestine, or brain) and excluded publications examining only the immune cell populations of affected patients. We recorded the number of hypo- or hypermethylated CpG sites and regions reported in each publication (from both main-text and supplementary data) that overlap with the ls-DMPs we have identified. Additionally, we assessed each publication on a number of criteria: (1) the tissue type studied as either bulk or sorted cells, (2) the methodology used to gather data, (3) the use of controls or cellular deconvolution techniques to mitigate the impact of multiple cells on data analysis, and (4) overall discussion of cellular heterogeneity, and in particular, leukocyte infiltration into tissues in samples and how it related to the study's findings.

Abbreviations

CpG: Cytosine phosphate guanine; cfDNA: Cell-free deoxyribonucleic acid; DMP: Differentially methylated position; ls-DMP: Leukocyte-specific differentially methylated position; CGI: CpG island; PBMC: Peripheral blood mononuclear cell; ROC: Receiver operating characteristic; TPR: True-positive rate; FPR: False-positive rate; IBD: Inflammatory bowel disease; AD: Alzheimer disease; TCGA: The Cancer Genome Atlas; FACS: Florescence activated cell sorting; PBS: Phosphate-buffered saline; IFG: Inferior frontal gyrus; MTG: Middle temporal gyrus; EC: Entorhinal cortex; IEC: Intestinal epithelial cells; PFC: Prefrontal cortex; STG: Superior temporal gyrus; UTR: Untranslated region; TSS: Transcription start site.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s43682-022-00011-z.

Additional file 1. Table S1: CpG sites with a Δ β-value of ≥ 0.8 between leukocytes and all other tissues within the Moss et al., 2018 dataset (GSE122126). **Table S2.** List of publications and relevant traits identified using the EWAS ATLAS that report one or more of our CpG sites from Table S1 as being differently methylated in their cohort. **Table S3.** List of publication titles and DOIs from publications listed in Table 2. **Table S4.** Table of publciations identifed from either the EWAS ATLAS or a PubMed search of a relavent trait that report one or more of our CpG sites from Table S1 as being differently methylated in their cohort. **Table S5.** Table of publciations identifed from either the EWAS ATLAS or a PubMed search of a relavent trait that report a differently methylated region containing one of our CpG sites from Table S1. **Table S6.** CpG sites with a Δ β-value of ≥ 0.8 between leukocytes and cortical neurons within the Moss et al., 2018 dataset (GSE122126). **Table S7.** List of primers used in this study.

Additional file 2: Figure S1. Distribution of ls-DMPs with respect to genes and CpG islands. Pie charts showing the proportion of hyper- (a - b) and hypo-methylated (c - d) ls-DMPs in relation to the closest gene and the nearest CpG island, respectively. **Figure S2.** Heatmap of whole genome bisulfite sequencing at the HOXA3 locus from 19 different mouse tissues (chr6: 52125343-52127779, mm9; human equivalent is chr7:27,113,454-27,115,864, hg38). Data was separated into 100 base pair windows and overall methylation was determined with SeqMonk. **Figure S3.** Deconvolution of in vitro mixed PBMC and Intestinal organoid DNA based on MAP4K1 DNA methylation. (a) Example methylation heatmap outputs of the MAP4K1 amplicon for PBMCs (left), 50:50 mix (center), and Intestinal organoid (right). Each row represents an individual read, and each column is a CpG site within the amplicon; column names refer to the CpG position in the read. (b) Density plot for the number of methylated CpG sites per read in pure Intestinal organoid (blue) and pure PBMC (red) samples. The y-axis represents the probability per unit on the x-axis such that the area under the curve for a specific interval is equal to the probability of the number of methylated CpGs in that interval. Bandwidth is 0.8 (c) Stacked bar chart of the proportion of classified for each sample type. Reads were classified as Intestinal organoid (blue), PBMC (red), or unclassified (grey). (d) Scatter plot of observed read classification vs expected read classification for Intestinal organoid and PBMC. **Figure S4.** All heatmap outputs for HOXA3 in vitro mixes. Row 1: Intestinal Organoid replicates 1-3. Row 2: Intestinal Organoid replicates 4-6. Row 3: in vitro mixes 99:1, 95:5, 50:50 (as PBMC DNA : Intestinal organoid DNA). Row 4: in vitro mixes 5:95, 1:99, PBMC replicate 1 (as PBMC DNA : Intestinal organoid DNA). Row 5: PBMC replicates 2-3. **Figure S5.** All heatmap outputs for MAP4K1 in vitro mixes. Row 1: Intestinal Organoid replicates 1-3. Row 2: Intestinal Organoid replicates 4-6. Row 3: in vitro mixes 99:1, 95:5, 50:50 (as PBMC DNA : Intestinal organoid DNA). Row 4: in vitro mixes 5:95, 1:99, PBMC replicate 1 (as PBMC DNA : Intestinal organoid DNA). Row 5: PBMC replicates 2-3. **Figure S6.** Binomial logistic regression for PBMC and Intestinal organoid in vitro mixes. Binomial logistic regression (a and c) and receiver operating characteristic (b and d) for the HOXA3 and MAP4K1 amplicons, respectively.

Additional file 3: Figure S7. Deconvolution of saliva based upon MAP4K1 DNA methylation patterns. (a) Example methylation heatmap outputs of the MAP4K1 amplicon for salivary leukocytes (left), raw saliva (center), and buccal epithelium (right). Each row represents an individual

read, and each column is a CpG site within the amplicon; column names refer to the CpG position in the read. (b) Density plot for the number of methylated CpG sites per read in pure buccal epithelium (blue) and pure salivary leukocytes (red) samples. The y-axis represents the probability per unit on the x-axis such that the area under the curve for a specific interval is equal to the probability of the number of methylated CpGs in that interval. Bandwidth is 0.45 (c) Scatter plot of observed read classification vs expected read classification for three mixed saliva technical replicates. Expected percentage based on manual cell counts. **Figure S8.** All heatmap outputs for HOXA3 saliva samples. Row 1: Salivary leukocytes 1.1 – 1.3, Salivary leukocytes 2.1. Row 2: Salivary leukocytes 2.2 – 2.3, Salivary leukocytes 3.1 – 3.2. Row 3: Salivary leukocytes 3.3, Raw saliva 1.1 – 1.3. Row 3: Buccal epithelium 1.1 – 1.3, Buccal epithelium 2.1. Row 3: Buccal epithelium 2.2 – 2.3, Buccal epithelium 3.1 – 3.2. Row 4: Buccal epithelium 3.3. **Figure S9.** All heatmap outputs for MAP4K1 saliva samples. Row 1: Salivary leukocytes 1.1 – 1.3, Salivary leukocytes 2.1. Row 2: Salivary leukocytes 2.2 – 2.3, Salivary leukocytes 3.1 – 3.2. Row 3: Salivary leukocytes 3.3, Raw saliva 1.1 – 1.3. Row 3: Buccal epithelium 1.1 – 1.3, Buccal epithelium 2.1. Row 3: Buccal epithelium 2.2 – 2.3, Buccal epithelium 3.1 – 3.2. Row 4: Buccal epithelium 3.3. **Figure S10.** Binomial logistic regression for saliva samples. Binomial logistic regression (a and c) and receiver operating characteristic (b and d) for the HOXA3 and MAP4K1 amplicons, respectively. **Figure S11.** Estimation of increasing leukocytes in the Alzheimer disease brain. (a) DNA methylation (expressed as a β-value) at cg00921266 (HOXA3) in Braak stage 0 and Braak stage 6 prefrontal cortices. (b) The estimated increase in leukocytes within the Braak stage 6 brain based upon the increase in DNA methylation at cg00921266. Alzheimer Disease methylation data obtained from Smith et al., 2018 (GSE80970). **Figure S12.** The relationship between cg00921266 methylation and the proportion of leukocytes in individual cancers. Scatterplots of cg00921266 DNA methylation (expressed as a β-value vs a leukocyte estimate for all individual cancers in the TCGA pan-cancer dataset. ACC = Adrenocortical carcinoma; BRCA = Breast invasive carcinoma; CESC = Cervical squamous cell carcinoma and endocervical adenocarcinoma; CHOL = Cholangiocarcinoma; COAD = Colon adenocarcinoma; ESCA = Esophageal carcinoma; GBM = Glioblastoma multiforme; HNSC = Head and Neck squamous cell carcinoma; KICH = Kidney Chromophobe; KIRC = Kidney renal clear cell carcinoma; KIRP = Kidney renal papillary cell carcinoma; LGG = Brain Lower Grade Glioma; LIHC = Liver hepatocellular carcinoma; LUAD = Lung adenocarcinoma; LUSC = Lung squamous cell carcinoma; PAAD = Pancreatic adenocarcinoma; PCPG = Pheochromocytoma and Paraganglioma; PRAD = Prostate adenocarcinoma; READ = Rectum adenocarcinoma; SARC = Sarcoma; SKCM = Skin Cutaneous Melanoma; STAD = Stomach adenocarcinoma; UCS = Uterine Carcinosarcoma; UCEC = Uterine Corpus Endometrial Carcinoma. Excludes thyroid carcinoma, bladder urothelial carcinoma, mesothelioma, testicular germ cell cancer, uveal melanoma, ovarian serous cystadenocarcinoma (found in Figure 5). Data was obtained from the TCGA Pan-cancer atlas publication page (https://gdc.cancer.gov/about-data/publications/pancanatlas).

Acknowledgements
We acknowledge A. Prof Keith Ooi in collection of the rectal biopsies for Molecular and Integrative Cystic Fibrosis Research Centre at Sydney Children's Hospital. SAW is supported by an Australian National Health and Medical Research Council grant (NHMRC_APP1188987). We thank Michelle Wilson for assistance with flow cytometry. We thank Dr. Donna Bond for technical assistance and for reviewing the manuscript.

Authors' contributions
RW performed the PBMC isolation and DNA purification. SW cultured intestinal organoids and performed the DNA purification. MD performed all other experimental work. OO provided extensive advice on statistical analysis. MD wrote the first draft of the manuscript and prepared the figures. MD and OO were responsible for custom R-scripts used in this study. TH and IM conceived the study. The authors read and approved the final manuscript.

Funding
This study was funded by the University of Otago.

Availability of data and materials
All data generated or analysed during this study are included in this published article [and its supplementary information files].

Declarations

Ethics approval and consent to participate
Collection and culture of intestinal organoid cells
Participants' provided written informed consent for rectal biopsy's and subsequent culture of intestinal organoids. The sample collection was approved by the Sydney Children's Hospital Ethics Review Board (HREC/16/SCHN/120).
PBMC and saliva collection
Peripheral blood mononuclear cells and saliva samples were collected from the researchers with informed consent.

Consent for publication
Not applicable.

Competing interests
TH is a shareholder and director of Totovision/Totogen Ltd, a small agricultural and biotechnology consultancy. The other authors declare no competing interests.

Author details
[1]Department of Anatomy, University of Otago, Dunedin, New Zealand. [2]School of Medical Sciences, Faculty of Medicine and Health, University of New South Wales, Sydney, Australia. [3]Molecular and Integrative Cystic Fibrosis Research Centre (miCF_RC), University of New South Wales and Sydney Children's Hospital, Randwick, Australia. [4]Department of Respiratory Medicine, Sydney Children's Hospital, Randwick, Australia. [5]Department of Pathology, University of Otago, Dunedin, New Zealand.

Received: 23 February 2022 Accepted: 6 May 2022
Published online: 23 June 2022

References

1. Goll MG, Bestor TH. Eukaryotic cytosine methyltransferases. Annu Rev Biochem. 2005;74:481–514.
2. Rao S, et al. Systematic prediction of DNA shape changes due to CpG methylation explains epigenetic effects on protein–DNA binding. Epigenetics Chromatin. 2018;11(1):6.
3. Du Q, et al. Methyl-CpG-binding domain proteins: readers of the epigenome. Epigenomics. 2015;7(6):1051–73.
4. Schmidt M, et al. Deconvolution of cellular subsets in human tissue based on targeted DNA methylation analysis at individual CpG sites. BMC Biol. 2020;18(1):178.
5. Hon GC, et al. Epigenetic memory at embryonic enhancers identified in DNA methylation maps from adult mouse tissues. Nat Genet. 2013;45(10):1198–206.
6. Bendich A, Wilczok T, Borenfreund E. Circulating DNA as a possible factor in oncogenesis. Science. 1965;148(3668):374–6.
7. Stroun M, et al. Neoplastic characteristics of the DNA found in the plasma of cancer patients. Oncology. 1989;46(5):318–22.
8. Lehmann-Werman R, et al. Monitoring liver damage using hepatocyte-specific methylation markers in cell-free circulating DNA. JCI Insight. 2018;3(12):e120474.
9. Lehmann-Werman R, et al. Identification of tissue-specific cell death using methylation patterns of circulating DNA. Proc Natl Acad Sci. 2016;113(13):E1826.
10. Moss J, et al. Comprehensive human cell-type methylation atlas reveals origins of circulating cell-free DNA in health and disease. Nat Commun. 2018;9(1):5068.
11. Reinius LE, et al. Differential DNA methylation in purified human blood cells: implications for cell lineage and studies on disease susceptibility. PLoS ONE. 2012;7(7):e41361.
12. You C, et al. A cell-type deconvolution meta-analysis of whole blood EWAS reveals lineage-specific smoking-associated DNA methylation changes. Nat Commun. 2020;11(1):4779.

13. Thorsson V, et al. The immune landscape of cancer. Immunity. 2018;48(4):812-830.e14.
14. Titus AJ, et al. Deconvolution of DNA methylation identifies differentially methylated gene regions on 1p36 across breast cancer subtypes. Sci Rep. 2017;7(1):11594.
15. Lessard S, et al. Comparison of DNA methylation profiles in human fetal and adult red blood cell progenitors. Genome Medicine. 2015;7(1):1.
16. Lange V, et al. Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. BMC Genomics. 2014;15(1):63.
17. Sato T, et al. Single Lgr5 stem cells build crypt-villus structures in vitro without a mesenchymal niche. Nature. 2009;459(7244):262–5.
18. Theda C, et al. Quantitation of the cellular content of saliva and buccal swab samples. Sci Rep. 2018;8(1):6944.
19. Nishitani S, et al. DNA methylation analysis from saliva samples for epidemiological studies. Epigenetics. 2018;13(4):352–62.
20. Li M, et al. EWAS Atlas: a curated knowledgebase of epigenome-wide association studies. Nucleic Acids Res. 2019;47(D1):D983-d988.
21. Chandra A, et al. Epigenome-wide DNA methylation regulates cardinal pathological features of psoriasis. Clin Epigenetics. 2018;10(1):108.
22. Agliata I, et al. The DNA methylome of inflammatory bowel disease (IBD) reflects intrinsic and extrinsic factors in intestinal mucosal cells. Epigenetics. 2020;15(10):1068–82.
23. Zhang L, et al. Epigenome-wide meta-analysis of DNA methylation differences in prefrontal cortex implicates the immune processes in Alzheimer's disease. Nat Commun. 2020;11(1):6114.
24. Haertle L, et al. Methylomic profiling in trisomy 21 identifies cognition- and Alzheimer's disease-related dysregulation. Clin Epigenetics. 2019;11(1):195.
25. Semick SA, et al. Integrated DNA methylation and gene expression profiling across multiple brain regions implicate novel genes in Alzheimer's disease. Acta Neuropathol. 2019;137(4):557–69.
26. Zhou F, et al. Epigenome-wide association analysis identified nine skin DNA methylation loci for psoriasis. J Investig Dermatol. 2016;136(4):779–87.
27. Verma D, et al. Genome-wide DNA methylation profiling identifies differential methylation in uninvolved psoriatic epidermis. J Investig Dermatol. 2018;138(5):1088–93.
28. Harris RA, et al. DNA methylation-associated colonic mucosal immune and defense responses in treatment-naïve pediatric ulcerative colitis. Epigenetics. 2014;9(8):1131–7.
29. Gasparoni G, et al. DNA methylation analysis on purified neurons and glia dissects age and Alzheimer's disease-specific changes in the human cortex. Epigenetics Chromatin. 2018;11(1):41.
30. Smith RG, et al. Elevated DNA methylation across a 48-kb region spanning the HOXA gene cluster is associated with Alzheimer's disease neuropathology. Alzheimers Dement. 2018;14(12):1580–8.
31. Li QS, Sun Y, Wang T. Epigenome-wide association study of Alzheimer's disease replicates 22 differentially methylated positions and 30 differentially methylated regions. Clin Epigenetics. 2020;12(1):149.
32. De Jager PL, et al. Alzheimer's disease: early alterations in brain DNA methylation at ANK1, BIN1, RHBDF2 and other loci. Nat Neurosci. 2014;17(9):1156–63.
33. Altuna M, et al. DNA methylation signature of human hippocampus in Alzheimer's disease is linked to neurogenesis. Clin Epigenetics. 2019;11(1):91.
34. Smith RG, et al. A meta-analysis of epigenome-wide association studies in Alzheimer's disease highlights novel differentially methylated loci across cortex. Nat Commun. 2021;12(1):3517.
35. Pietronigro E, Zenaro E, Constantin G. Imaging of leukocyte trafficking in Alzheimer's disease. Front Immunol. 2016;7(33).
36. Gemechu JM, Bentivoglio M. T cell recruitment in the brain during normal aging. Front Cell Neurosci. 2012;6:38.
37. Coussens LM, Werb Z. Inflammation and cancer. Nature. 2002;420(6917):860–7.
38. Di Vinci A, et al. Quantitative methylation analysis of HOXA3, 7, 9, and 10 genes in glioma: association with tumor WHO grade and clinical outcome. J Cancer Res Clin Oncol. 2012;138(1):35–47.
39. Daugaard I, et al. Identification and validation of candidate epigenetic biomarkers in lung adenocarcinoma. Sci Rep. 2016;6:35807.
40. Kuasne H, et al. Genome-wide methylation and transcriptome analysis in penile carcinoma: uncovering new molecular markers. Clin Epigenetics. 2015;7(1):46.
41. Gan BL, et al. Downregulation of HOXA3 in lung adenocarcinoma and its relevant molecular mechanism analysed by RT-qPCR, TCGA and in silico analysis. Int J Oncol. 2018;53(4):1557–79.
42. Kleiveland CR, et al. Peripheral blood mononuclear cells. In: Verhoeckx K, et al., editors. The Impact of Food Bioactives on Health: in vitro and ex vivo models. Cham: Springer International Publishing; 2015. p. 161–7.
43. Alharbi RA, et al. The role of HOX genes in normal hematopoiesis and acute leukemia. Leukemia. 2013;27(5):1000–8.
44. Ramos-Mejía V, et al. HOXA9 promotes hematopoietic commitment of human embryonic stem cells. Blood. 2014;124(20):3065–75.
45. Sauvageau G, et al. Differential expression of homeobox genes in functionally distinct CD34+ subpopulations of human bone marrow cells. Proc Natl Acad Sci U S A. 1994;91(25):12223–7.
46. Ottersbach K. Endothelial-to-haematopoietic transition: an update on the process of making blood. Biochem Soc Trans. 2019;47(2):591–601.
47. Dou DR, et al. Medial HOXA genes demarcate haematopoietic stem cell fate during human development. Nat Cell Biol. 2016;18(6):595–606.
48. Magnusson M, et al. HOXA10 is a critical regulator for hematopoietic stem cells and erythroid/megakaryocyte development. Blood. 2007;109(9):3687–96.
49. Iacovino M, et al. HoxA3 is an apical regulator of haemogenic endothelium. Nat Cell Biol. 2011;13(1):72–8.
50. Chuang HC, Wang X, Tan TH. MAP4K family kinases in immunity and inflammation. Adv Immunol. 2016;129:277–314.
51. Jeziorska DM, et al. DNA methylation of intragenic CpG islands depends on their transcriptional activity during differentiation and disease. Proc Natl Acad Sci. 2017;114(36):E7526.
52. Uhlén M, et al. Proteomics Tissue-based map of the human proteome. Science. 2015;347(6220):1260419.
53. Langie SAS, et al. Salivary DNA methylation profiling: aspects to consider for biomarker identification. Basic Clin Pharmacol Toxicol. 2017;121(Suppl Suppl 3):93–101.
54. Wong YT, et al. A comparison of epithelial cell content of oral samples estimated using cytology and DNA methylation. Epigenetics. 2022;17(3):327–34.
55. Bjarnason I. The use of fecal calprotectin in inflammatory bowel disease. Gastroenterology Hepatology. 2017;13(1):53–6.
56. Ayling RM, Kok K. Fecal calprotectin. Adv Clin Chem. 2018;87:161–90.
57. Dagogo-Jack I, Shaw AT. Tumour heterogeneity and resistance to cancer therapies. Nat Rev Clin Oncol. 2018;15(2):81–94.
58. Guo M, et al. Epigenetic heterogeneity in cancer. Biomarker Res. 2019;7(1):23.
59. Gonzalez H, Hagerling C, Werb Z. Roles of the immune system in cancer: from tumor initiation to metastatic progression. Genes Dev. 2018;32(19–20):1267–84.
60. Hoadley KA, et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173(2):291-304.e6.
61. Wong SL, et al. Molecular dynamics and theratyping in airway and gut organoids reveal R352Q-CFTR conductance defect. bioRxiv. 2021. p. 2021.08.11.456003.
62. Berkers G, et al. Rectal organoids enable personalized treatment of cystic fibrosis. Cell Rep. 2019;26(7):1701-1708.e3.
63. Oberacker P, et al. Bio-On-Magnetic-Beads (BOMB): Open platform for high-throughput nucleic acid extraction and manipulation. PLoS Biol. 2019;17(1):e3000107.
64. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):10–2.
65. Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. Bioinformatics. 2011;27(11):1571–2.