Inferring gene regulatory networks using DNA methylation data

Thomas E. Bartlett[1*], Melodie Li[1], and Qiulin Huang[2,3]

[1]Department of Statistical Science, University College London, London, U.K.
[2]Human Embryo and Stem Cell Laboratory, The Francis Crick Institute, London, U.K.
[3]Department of Physiology, Development and Neuroscience, The Centre for Trophoblast Research, University of Cambridge, Cambridge, U.K.
[*]thomas.bartlett.10@ucl.ac.uk

Abstract

We show much-improved accuracy of inference of GRN (gene regulatory network) structure, resulting from the use of an epigenomic prior network. We also find that DNAme data are very effective for inferring the epigenomic prior network, recapitulating known epigenomic network structure found previously from chromatin accessibility data, and typically providing potential TF cis-regulations for at least eight times as many genes when compared with chromatin accessibility data. When our proposed methodology is applied to real datasets from human embryonic development and from women at risk of breast cancer, we find patterns of differential cis-regulation that are in line with expectations under appropriate biological models, and that can be used to identify pre-cancerous epigenomic changes with valid functional genomic interpretations.

Keywords

DNA methylation; gene regulatory networks; human embryonic development; cancer prediction.

1 Introduction

Gene-regulatory network (GRN) models have been shown to be effective at finding specific panels of transcription factors (TFs) that are associated with cell type identities and with progression of certain diseases (see for example [1–3]). These successful methods of GRN inference are typically based on the target-gene approach [4–6] (Equations 1 and 2), which can find important structure in 'omic networks such as feedback loops [7]. Here, we propose target-gene GRN inference methodology that builds on our previous work to infer networks from epigenomic data such as DNA methylation (DNAme) data [8, 9], and using advanced regression methods with scRNA-seq (single-cell RNA-seq) and multimodal data [10–12]. The methodology that we propose here infers the GRN using advanced regression methods with scRNA-seq (gene-expression) data following estimation of an epigenomic prior network. The sequence of these operations is crucial: we show that the accuracy of GRN inference from gene expression data is much improved when using an epigenomic prior network that has been inferred in advance. This contrasts with alternative strategies, in which the GRN is estimated by combining networks inferred independently in parallel from gene expression data and from chromatin accessibility data [1, 2, 5]. Importantly, with our proposed methodology the gene-expression data can originate from a completely independent source to the data that is used to infer the epigenomic prior network. This enables novel analyses specific to tissue and cell-type, by making use of the wealth of genomic and epigenomic data that is now publicly available [13].

A gene regulatory network (GRN) represents all the gene targets of every TF in a specific cell type. Hence, GRNs are important summaries of the genome-wide gene regulation profile that characterises a particular cellular identity. GRNs have been inferred successfully using methods based on random forests (such as GENIE3 / GRNboost2 [5, 6] as part of the SCENIC / SCENIC+ pipelines [1–3]), as well as alternative approaches based on mutual information [12, 17–20] and recent extensions [21]. However, a genomic network that has been inferred only from gene expression data (such as scRNA-seq data) is more accurately referred to as a gene co-expression network, unless data is included on the interactions between gene-products and DNA [10, 22, 23]. Hence, in order to infer gene regulatory networks, data must be included in the inference procedure that allows more precise inference about the interaction of the gene product of the regulating gene (i.e., a TF) with the DNA near the regulated gene (cis-regulation) [1, 3, 11, 12]. Such data often combine the DNA location of cis-regulatory TF binding sites with epigenomic data corresponding to those locations such as chromatin accessibility data e.g. ATAC-seq. However, alternative epigenomic data-sources include DNAme data, and ChIP-seq data for histone modifications such as H3K27ac, H3K4me3, and K3K27me3. In the target-gene approach to gene-regulatory network inference [4–6], an advanced regression model is used to model the response x_i (the expression-level of gene i, Figure 1), as:

$$x_i = f_i(\mathbf{x}_{\mathbf{PA}^{\text{prior}}(i)}, e_i), \tag{1}$$

where $\mathbf{PA}^{\text{prior}}(i)$ are the possible regulators of gene i, $\mathbf{x}_{\mathbf{PA}^{\text{prior}}(i)}$ is the vector of expression levels of these potential regulators, f_i is a (possibly non-linear) function, and e_i is an error or noise term. In our proposed methodology, the epigenomic prior network consists of the potential regulators of all p genes genome-wide, $\mathbf{PA}^{\text{prior}}(i)$, $i \in \{1, ..., p\}$, that are inferred from epigenomic data in combination with TF motif-DNA sequence matching (Figure 1). For a linear regression model, Eq.1 would become:

$$x_i = \mathbf{x}^{\top}_{\mathbf{PA}^{\text{prior}}(i)} \mathbf{b}(i) + e_i. \tag{2}$$

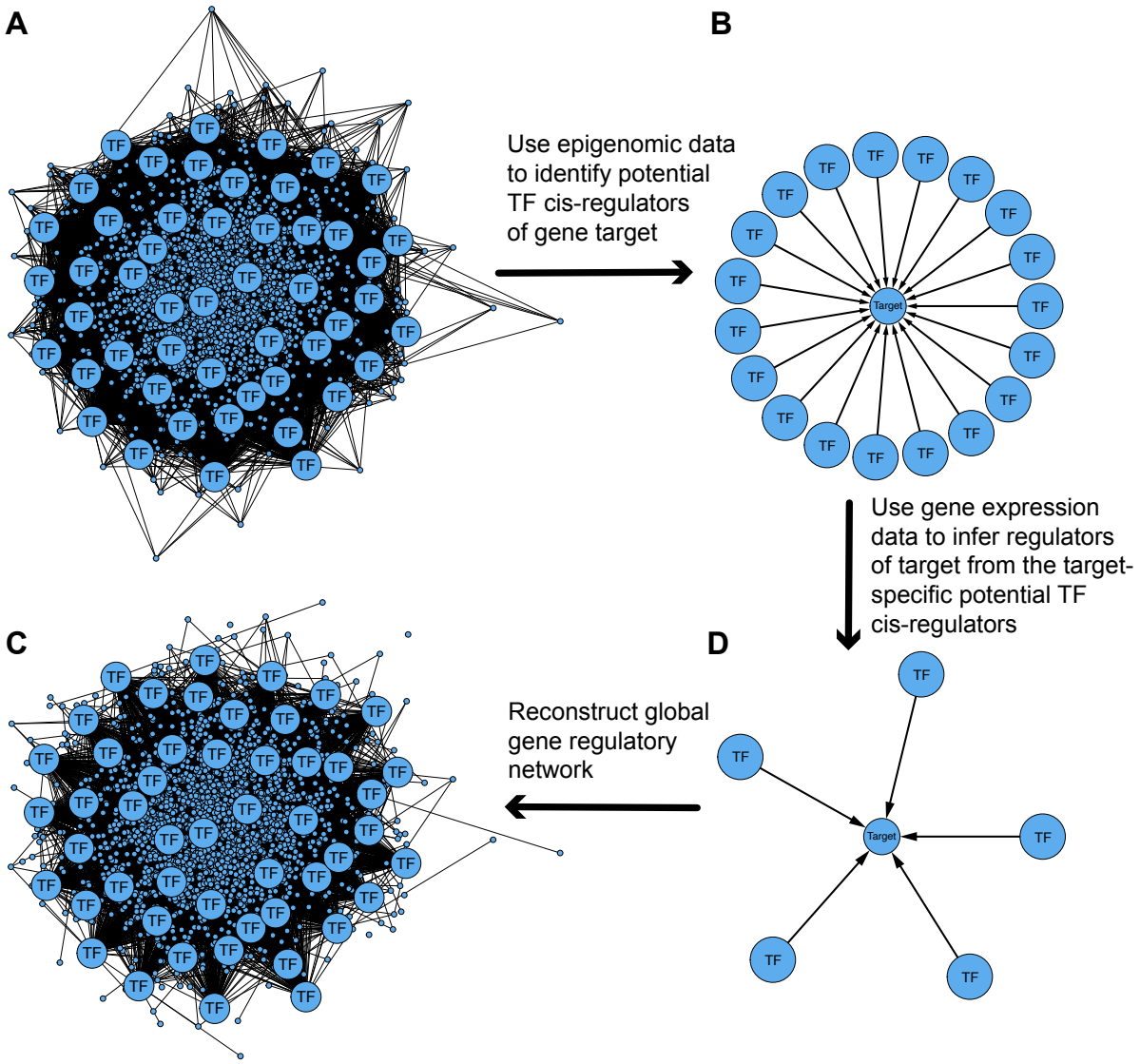

Figure 1: *(a) Epigenomic data (ATAC-seq / LiCAT-seq, H3K27ac, DNAme, etc) is used to infer the prior network structure, by analysing epigenomic patterns at cis-regulatory TF binding sites [14,15]. (b) There is improved accuracy in subsequent gene regulatory network inference from gene expression data. This is achieved because the prior epigenomic network narrows down the list of potential regulators of a target-gene from the order of 100s to 10s, which allows the advanced regression method to more accurately infer individual TF-gene cis-regulations (assessed by AUC, Fig 2). (c) Advanced regression methods such as random-forests (as used in GENIE3 /GRNboost2 [5, 6, 16]) are used to infer the cis-regulators of the target-gene from the prior selection provided by the epigenomic network. (d) The global gene cis-regulatory network is constructed, by combining the local networks that were inferred around each target-gene [4].*

In our proposed methodology, we first estimate the set of potential TF regulators $\mathbf{PA}^{\text{prior}}(i)$ separately for each target-gene i. This is part of a two-step inference procedure where the second step is feature selection for each i from the choice provided by $\mathbf{PA}^{\text{prior}}(i)$ using an advanced regression method (Eq. 1).

It is well known that changes in DNAme patterns precede cancer formation [24, 25]: these pre-cancerous changes are sometimes called 'epigenomic field-defects' [24, 26] or 'epigenetic field cancerisation' [27]. We have previously found that that unique epigenomic patterns distinguish cancer [28], and can be used to predict clinical outcomes [29], and it is well established that DNA methylation patterns enable different tissues and cell types to be distinguished from one another [30]. Changes in DNAme rates at regulatory DNA regions enable changes in gene regulation [31] and therefore allow changes of phenotype [32], motivating the methodology we propose here to systematically analyse genome-wide DNAme rates at putative binding sites of specific TFs. Recently, it has become clear that a large amount of the observed variation in DNAme patterns in bulk-tissue genomic data can be explained by the variation in the proportion of cells from different lineages that comprise the tissue sample [33]. Hence bulk-tissue DNAme data contains important information about the cell type composition of the tissue, and this finding can also be used as a clinical biomarker, by tracking cells-of-origin of cancer [27, 34]. In the methodology proposed here, we infer GRNs using DNAme data from specific cell types (rather than from a mixture of cells of different lineages), to allow inferences about the functional effects of DNAme changes at cis-regulatory TF binding sites. We use these cell type specific DNAme data to infer genomic networks, showing how DNAme data can be used to infer the structure of the prior epigenomic network. The prior epigenomic network is essential to achieve the most accurate inference from the advanced regression methods applied to gene expression data such as scRNA-seq data (Equation 1 and Figures 1 and 2). We show that the prior epigenomic networks we infer recapitulate well-validated TF cis-regulations of gene targets obtained using the widely used and well validated software packages pycistarget [1, 3, 15] and RGT-HINT [35]. We then infer gene-regulatory networks from gene-expression (scRNA-seq) data in combination with these epigenomic prior networks, in the contexts of a human embryonic development dataset, and a breast cancer at-risk dataset. We find gene regulatory networks characteristic of different lineage choices and disease at-risk states, with patterns of differential gene-expression that are in line with expectations under appropriate biological models.

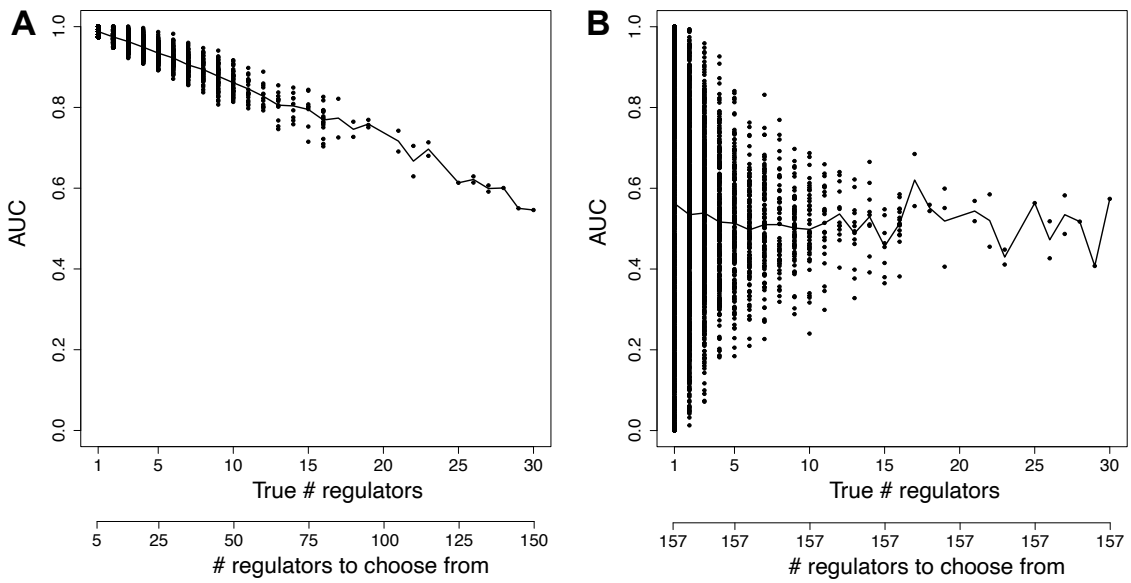

Figure 2: *(a) Random-forests target-gene GRN inference (Eq.1), implemented with xgboost and using prior network structure. (b) Random-forests target-gene GRN inference, implemented with GENIE3.*

2 Results

2.1 Using prior network structure from epigenomic data improves performance of regulatory network inference from gene expression data

We found that the epigenomic prior network allows the advanced regression method (Eq.1) to make more accurate predictions (Figure 2) of TF-gene cis-regulations (in comparison with known network structure). Figure 2 shows that there is much greater accuracy (assessed by AUC) in identifying known gene regulations when there are fewer potential regulators (of a target-gene). The results of Figure 2 are based on data generated from known network structure, that was simulated using realistic differential-equations models of chemical reaction kinetics [36] (Section 4). It can be seen from Figure 2 that there is an increase in accuracy of GRN inference resulting from the reduction in the number of potential TF regulators that the advanced regression method must choose from, from the order of 100s or 1000s down to the order of 10s. This reduction in number of potential regulators is achieved in our GRN inference procedure by the epigenomic prior network.

For each target-gene, the epigenomic prior network provides a list of possible TF regulators of the

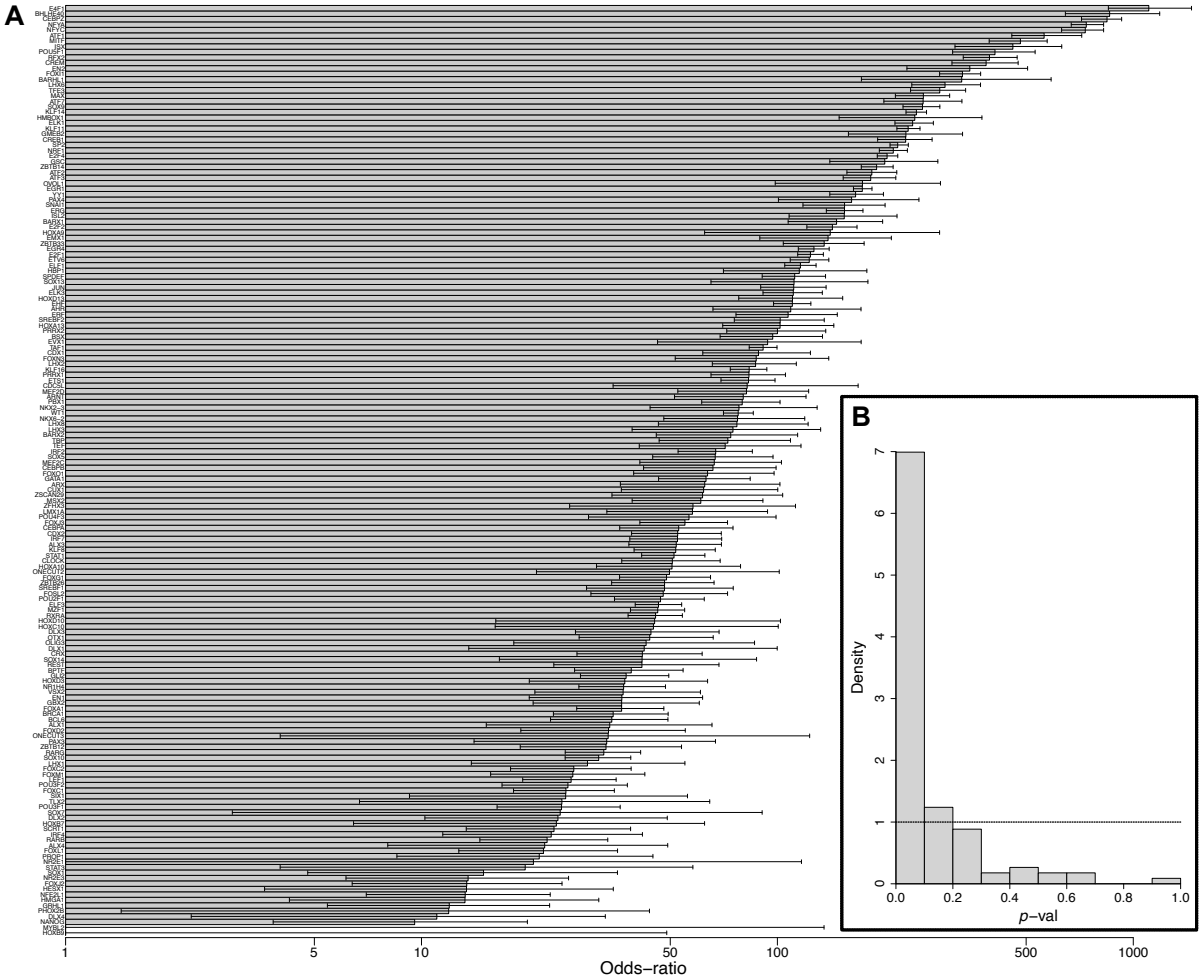

Figure 3: *(a) Tests of overlap of the DNA regions inferred as regulated by 170 TFs in the inner cell mass (ICM) according to both the modified-lever-cistarget (MLC) method and the RGT-HINT method [35] (as published previously [12]). Bars show odds-ratios (on a logarithmic scale), with whiskers showing 95% C.I.s. The corresponding significance p-vals (Fisher's exact test) are provided in Table S1. (b) The 124 motifs inferred when the pycistarget software is applied to the same LiCAT-seq data are are tested for significance according to the MLC method (z-test). The horizontal dashed line at density of 1 indicates the expected density for a uniform distribution of p-values, which would be expected if there were no correspondence between the TF-motif inferences from pycistarget and MLC for this LiCAT-seq dataset. LiCAT-seq is low-input chromatin accessibility data.*

target-gene that are specific to that target-gene. This target-gene specific list of potential regulators is typically a few 10s of TFs, making it much easier for the advanced regression method to identify the correct TF regulators of the target-gene (Figure 2a), than it would be from a much longer list of potential TFs, as required by GENIE3 / GRNboost2 and similar [1–3,5,6]. With existing methods such as GENIE3 / GRNboost2, the full list of TFs (typically of the order of 100s or 1000s of TFs) must be specified as potential regulators of each target-gene, meaning that it is much more challenging for those methods to find the true TF regulators of the target-gene from this much longer list of potential TFs that is not tailored specifically for the target-gene (Figure 2b). Hence, the epigenomic prior network allows greater accuracy of GRN inference from advanced regression methods such as xgboost [16], when used with single-cell gene expression data in a target-gene GRN inference procedure [4].

2.2 Recapitulating known epigenomic network structure using DNA methylation data

For the human embryonic development LiCAT-seq (low-input chromatin accessibility) dataset [37], we first verified whether TF-motif cis-regulations that were found using RGT-HINT [35] in previous work [12] would also be discovered from the same LiCAT-seq data by our proposed modified-lever-cistarget (MLC) method for epigenomic prior network inference (Section 4). We found 170 TFs that were present in both analyses, and found that the corresponding regulatory regions overlapped significantly for 168 out of these 170 TFs (Figure 3a). We also applied pycistarget [1, 3, 15] to the same LiCAT-seq human embryonic development dataset [37], finding 124 signficant TF motifs: a large number of these 124 motifs found by pycistarget were also found to have significant *p*-values (*z*-test) according to the MLC method (Figure 3b). Hence, we were able to validate the TF-motif cis-regulatory inferences found by MLC from the LiCAT-seq human embryonic development dataset [37], in comparison with TF-motif cis-regulations that were found from same dataset using RGT-HINT [35] in previous work [12], and also in comparison with inferences from the pycistarget method which is part of the SCENIC+ pipeline [3].

Applying the MLC method to the LiCAT-seq human embryonic development dataset [37] allows an epigenomic prior network to be inferred. Equivalently, applying our proposed MethCorTarget methodology (Section 4) to the DNAme human embryonic development dataset [38] allows an epigenomic prior network to be inferred. These epigenomic prior networks inferred from different epigenomic data sources are expected to contain overlapping and similar information. We verified this by testing the overlap between the TF motif-gene cis-regulations inferred by MLC as applied to the LiCAT-seq human embryonic development dataset [37], and the cis-regulations inferred by MethCorTarget applied to the DNAme human embryonic development dataset [38]: we found a highly significant overlap, odds-ratio $\mathrm{OR} = 13.3\ \ (12.9 - 13.8)$ ($p < 0.001$, Fisher's exact test). This highly significant overlap between the lists of cis-regulations inferred from the different data-sources illustrates the interchangeability between the epigenomic prior networks obtained for the same cell type from DNAme data and from chromatin accessibility data. We also note that epigenomic prior network structure inferred from DNAme may be more useful in certain applications, because it provides information about the regulation of more than eight times as many genes (23575 regulated genes found by MethCorTarget from DNAme data, compared to 2650 for MLC from LiCAT-seq data).

2.3 Inferring GRNs in epiblast and primitive endoderm cells of the human embryo

Epiblast (Epi) and primitive endoderm (PE, or hypoblast) cells are the main cell types of the inner cell mass (ICM) of the blastocyst of the human embryo at 5-7dpf (days post fertilisation) [39]. We have previously used ICM chromatin accessibility data to inform inference of GRNs from scRNA-seq data in Epi and PE [12]. We found that our epigenomic prior network obtained from DNAme data overlaps strongly with that obtained from chromatin accessibility data, but covers a much larger number of target-genes. We therefore proceeded to use the epigenomic prior network obtained from DNAme data as a valid epigenomic prior network to improve GRN inference from the human embryonic development scRNA-seq dataset [40]. We inferred GRNs from scRNA-seq data as described (Figure 1 and Section 4) for both Epi and PE cell types, using the scRNA-seq dataset cell type definitions shown in Figure S1 and Figure S2, and using the epigenomic prior network obtained from the DNAme human embryonic development dataset (ICM cells) [38].

An example of potential cis-regulations found as part of the GRN for Epi and PE is shown in Figure 4. The GRN was inferred by filtering potential cis-regulations found as part of the epigenomic prior network from DNAme data for ICM cells, by variable selection using advanced regression methods and scRNA-seq data for Epi and PE (according to the procedure outlined in Fig.1). Figure 4 shows DNAme

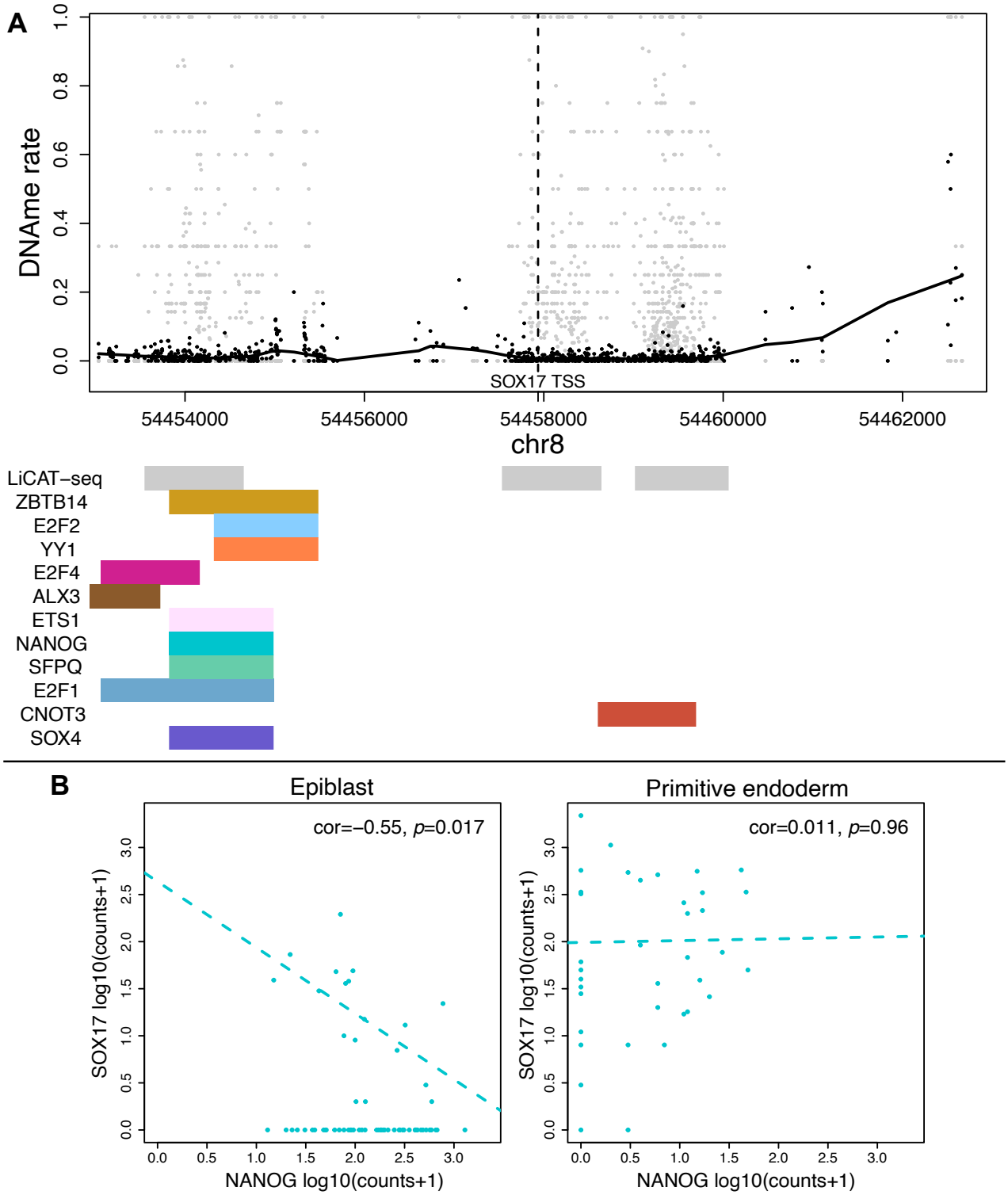

Figure 4: *(a) The DNAme profile of the cis-regulatory region (+/- 5kb) around the TSS (transcriptional start site) of the SOX17 gene (used as a marker for primitive endoderm cells in Figure S1, Figure S2, and Figure S3). Also shown are regions corresponding to previously-published and validated LiCAT-seq peaks [12,37], and regulatory regions corresponding to motifs of potential TF regulators inferred as part of the GRN. Mean DNAme rates are shown in black, and those for individual cells are shown in grey; non-linear trend-line is calculated with LOESS regression. (b) Scatter plots of log expression levels of SOX17 and one of its regulators NANOG [41] in Epi and PE.*

rate profiles for the 103 ICM cells (including epiblast and primitive endoderm, PE cells) in the DNAme human embryonic development dataset [38] for the *SOX17* gene, which we also used as a marker gene of primitive endoderm cells in Figure S1, Figure S2, and Figure S3. In Figure 4, changes in the DNAme profile are visible in regions corresponding to previously-published and validated LiCAT-seq peaks [12, 37], as well as around regulatory regions corresponding to the motifs of potential TF-regulators. Figure 4b also shows how in epiblast cells, the epiblast marker gene *NANOG* (Figure S1 and Figure S3) negatively regulates the PE marker gene *SOX17* (Figure S2 and Figure S3), which is consistent with *NANOG* maintaining Epi identity by repressing PE identity [41]. Also of interest, Figure 4 illustrates how the DNAme rates at these regulatory regions are highly variable, a feature we note is also of interest in understanding cancer formation and cancer predisposition [28, 29]. There are well established similarities between gene regulatory processes that have been observed in embryonic development and in cancer [42], and we hypothesise that phenotypic plasticity is represented by the variability in methylation rates that is visible in Figure 4 at CpG loci around the motifs of the potential TF cis-regulators of *SOX17*.

2.4 Differential regulation in epiblast and primitive endoderm cells

Differentially regulating TFs were identified as follows. For each TF, the differences in regulatory strength (measured by xgboost importance) were tested in Epi compared to PE, for all genes the TF can regulate according to the epigenomic prior network. Figure 5 shows that the differential regulation in Epi and PE is characterised by consistent upregulation of genes in PE, whereas in Epi there is significantly more downregulation of genes ($p < 0.001$, two-sample z-test of proportions). This follows from the differential expression analysis (comparing Epi with PE, Section 4), where the majority of genes were found to be upregulated in PE, relative to Epi. We note that this is a feature of differential gene regulation that we would expect to find if the epiblast can be thought of as the default cell type at the lineage bifurcation of the epiblast and hypoblast (primitive endoderm), as has been shown recently in human embyros [43]. In this case, specification of the hypoblast / primitive endoderm should be characterised by the activation in PE of a characteristic gene-regulatory network relative to Epi, such as the differential regulation network shown in Figure 5. We also tested enrichment of the TF targets of

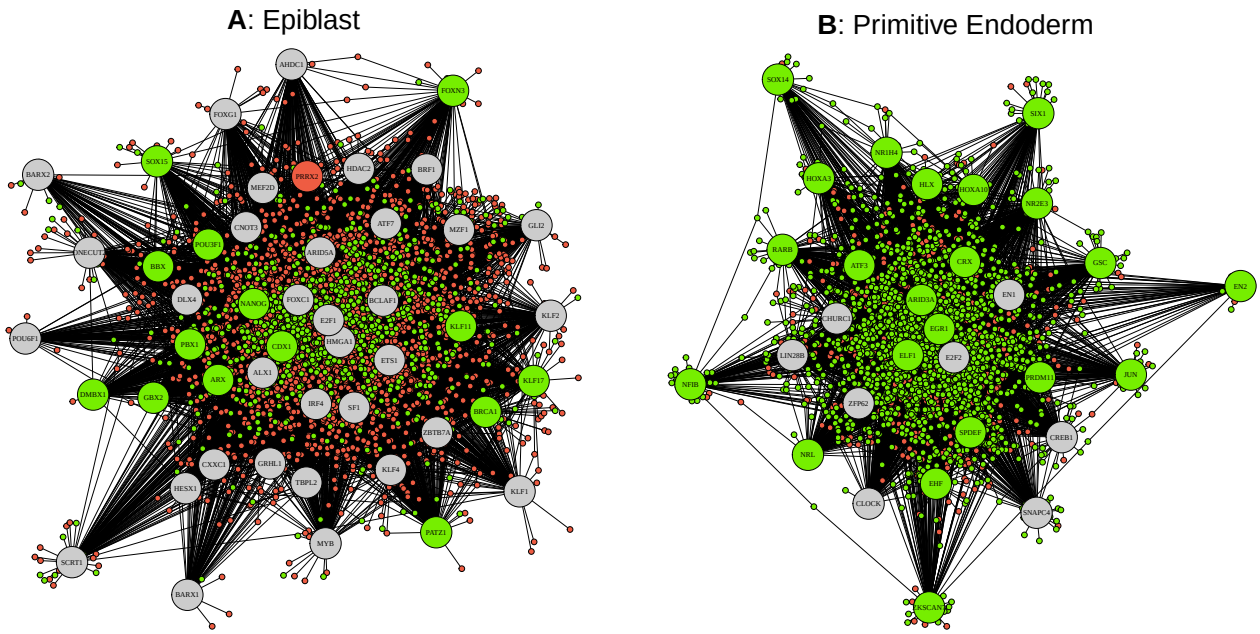

Figure 5: *(a) Epiblast (Epi) and (b) primitive endoderm (PE) differential regulation networks. Large circles indicate significantly differentially regulating transcription factors (Table S2); green and red indicate (respectively) significantly upregulated and downregulated TFs and genes; significance level is FDR p-val* < 0.05*, (t-test, Benjamini-Hochberg adjustment).*

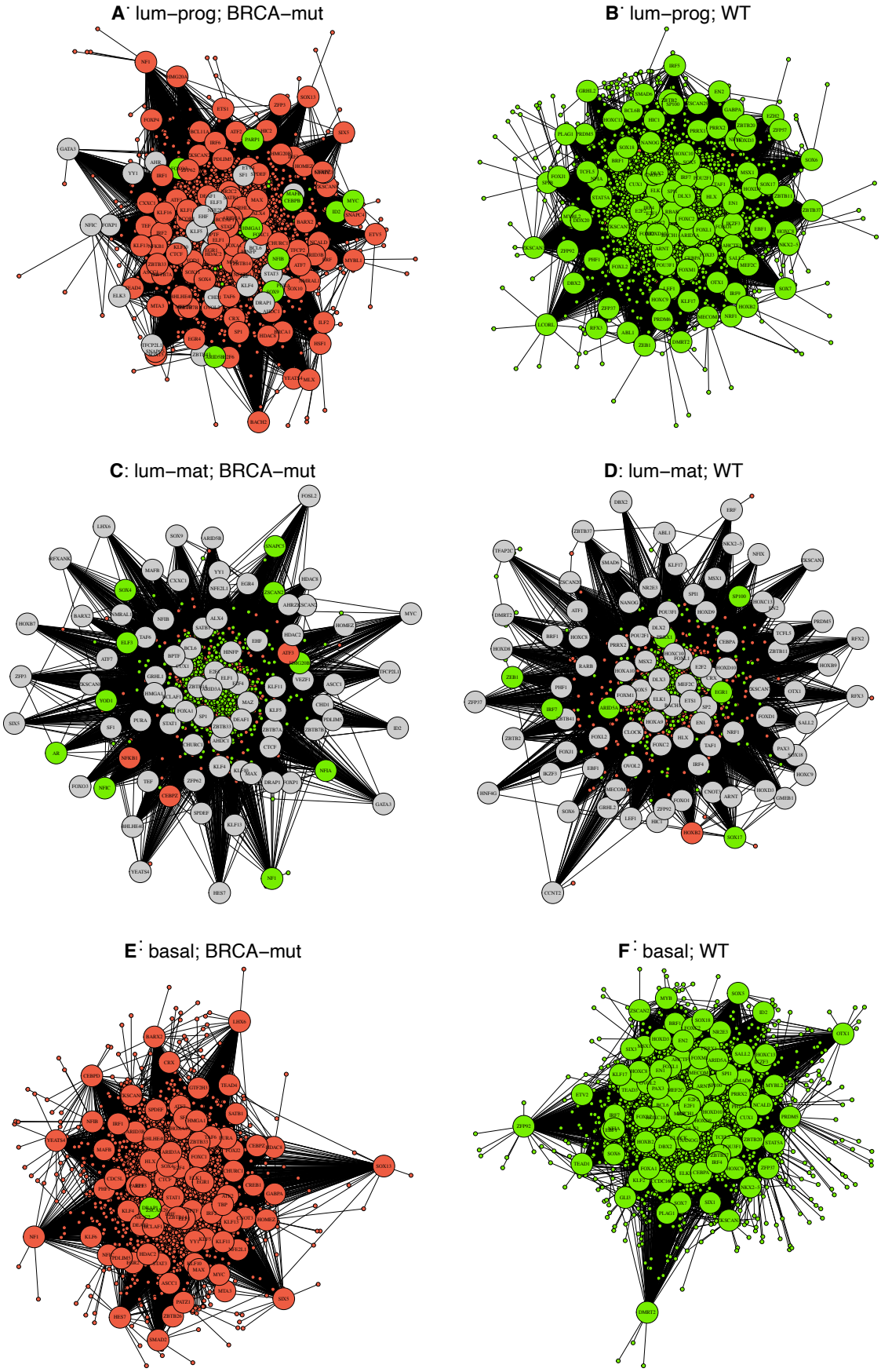

Figure 6: *Differential regulation networks in (a) and (b) luminal progenitor cells (lum-prog), in (c) and (d) mature luminal cells (lum-mat), and in (e) and (f) basal cells, for (a), (c) and (e) BRCA1 (FANCS) mutation carriers (BRCAmut), and (b), (d), and (f) BRCA1 (FANCS) wild-type (WT). Large circles indicate significantly differentially regulating transcription factors (Table S3); green and red indicate (respectively) significantly upregulated and downregulated TFs and genes; significance level is FDR* $p\text{-}val < 0.05$*, (t-test, Benjamini-Hochberg adjustment).*

the differentially regulating transcription factors shown in Figure 5, by comparison with 713 epiblast gene-sets defined and validated previously by domain-experts [12]. We confirmed that there was significantly more enrichment of these epiblast gene-sets amongst TF targets differentially regulated in epiblast compared to primitive endoderm ($p < 0.001$, Wilcoxon signed-rank test), i.e., when comparing the enrichment of TF targets in Fig.5a with the enrichment of those in Fig.5b.

2.5 Differential regulation in cells-of-origin of breast cancer

Women who are carriers of a mutation in the *BRCA1* (*FANCS*) gene have a greatly increased chance of suffering from breast cancer during their lifetime [44]. Hence, apparently healthy cells from *BRCA1* mutation carriers are likely to harbour pre-cancerous genomic changes, and insights can be gained into breast cancer formation and initiation by comparing healthy cells from *BRCA1* mutation carriers (BRCAmut) with cells from *BRCA1* wild-type (WT). Luminal progenitor, mature luminal, and basal cells are epithelial cell subtypes that are thought to have important roles in the formation and initiation of human breast cancer [45], and we have previously developed and validated DNAme cancer biomarkers based on definitions of these cell types [27, 34]. We inferred epigenomic prior networks for each of the cell types luminal progenitor, mature luminal, and basal, using MethCorTarget with the breast cancer at-risk DNAme dataset [46], and using MLC with the breast cancer at-risk H3K27ac ChIP-seq dataset [46]. We chose to validate the DNAme epigenomic prior network against an epigenomic prior network inferred from H3K27ac ChIP-seq data, because H3K27ac is a histone modification that is associated with activated gene-regulatory regions of the DNA [47, 48]. We found a highly significant overlap between the TF-gene regulations inferred by MLC applied to the H3K27ac data and the regulations inferred by MethCorTarget applied to the DNAme data in all cell types, hence validating the inferred DNAme epigenomic prior networks. The odds-ratios for the overlap between the gene regulations inferred from H3K27ac data and DNAme data according to these methods are OR $= 45.5$ $(44.8 - 46.2)$, OR $= 45.7$ $(45.0 - 46.5)$, and OR $= 43.2$ $(42.4 - 44.0)$, in luminal progenitor, mature luminal, and basal cells (respectively); statistical significance is $p < 0.001$ (Fisher's exact test) for all cell types. We also note that, interestingly, the epigenomic prior networks for luminal progenitor, mature luminal, and basal cells inferred from DNAme data again show potential TF cis-regulations for at least eight time as many genes (respectively, 34792, 34680, and 34021 genes) than for chromatin accessibility (H3K27ac) data (respectively 4222, 4194, and 3597 different genes).

Using the cell type specific DNAme epigenomic prior networks for luminal progenitor, mature luminal, and basal cell types, we inferred GRNs as described (Section 4) from the scRNA-seq breast cancer at-risk dataset [50], based on the definitions of these cell types shown in Figure S4. We inferred GRNs in this way for scRNA-seq data from cells both from BRCAmut and WT, using the epigenomic prior networks obtained from the DNAme breast cancer at-risk dataset [46]. Differentially regulating TFs were then identified as follows. For each TF, the differences in regulatory strength (measured by xgboost importance) were tested in BRCAmut compared to WT, for all genes the TF can regulate according to the epigenomic prior network. Figure 6a, b, e, and f show that in luminal progenitor and in basal cells, the significant differential regulation in *BRCA1* (*FANCS*) mutation carriers (BRCAmut) is overwhelmingly characterised by downregulation of transcription factors and in the genes that they regulate compared to upregulation in *BRCA1* wild-type (WT) ($p < 0.001$ for both these cell types, two-sample z-test of proportions), reflecting the fact that in the differential expression analysis (comparing BRCAmut with WT, Section 4), the majority of genes were found to be upregulated in WT. However, the equivalent effect in mature luminal cells (Figure 6a and b) is small, corresponding to slightly more upregulation in BRCAmut compared to WT. We hypothesise that this finding means that in BRCAmut, non-terminally differentiated cells such as luminal progenitor and basal retain more flexibility over cell-fate choices than in WT, which contributes to their carcinogenic potential.

We also used gene-set enrichment analysis (GSEA) to further investigate the association with cancer of the genes regulated by the differentially regulating (diff-reg) TFs shown in Figure 6. We found that there are a significant number of Broad Institute Molecular Signatures Database (MSigDB) cancer gene-sets [49] that are enriched by the gene targets of the diff-reg TFs inferred in BRCAmut vs. WT in all cell types (Figure 7a). However, amongst these BRCAmut vs. WT diff-reg TF targets, the number of significantly enriched cancer gene-sets is much greater for luminal progenitor and basal cells, consistent with these cell types being cells-of-origin of aggressive cancers, and with *BRCA* mutation carriers harbouring pre-cancerous changes to gene regulation [27, 45]. Next, we reasoned that the genes that are co-regulated by the greatest number of significantly diff-reg TFs in *BRCA* mutation

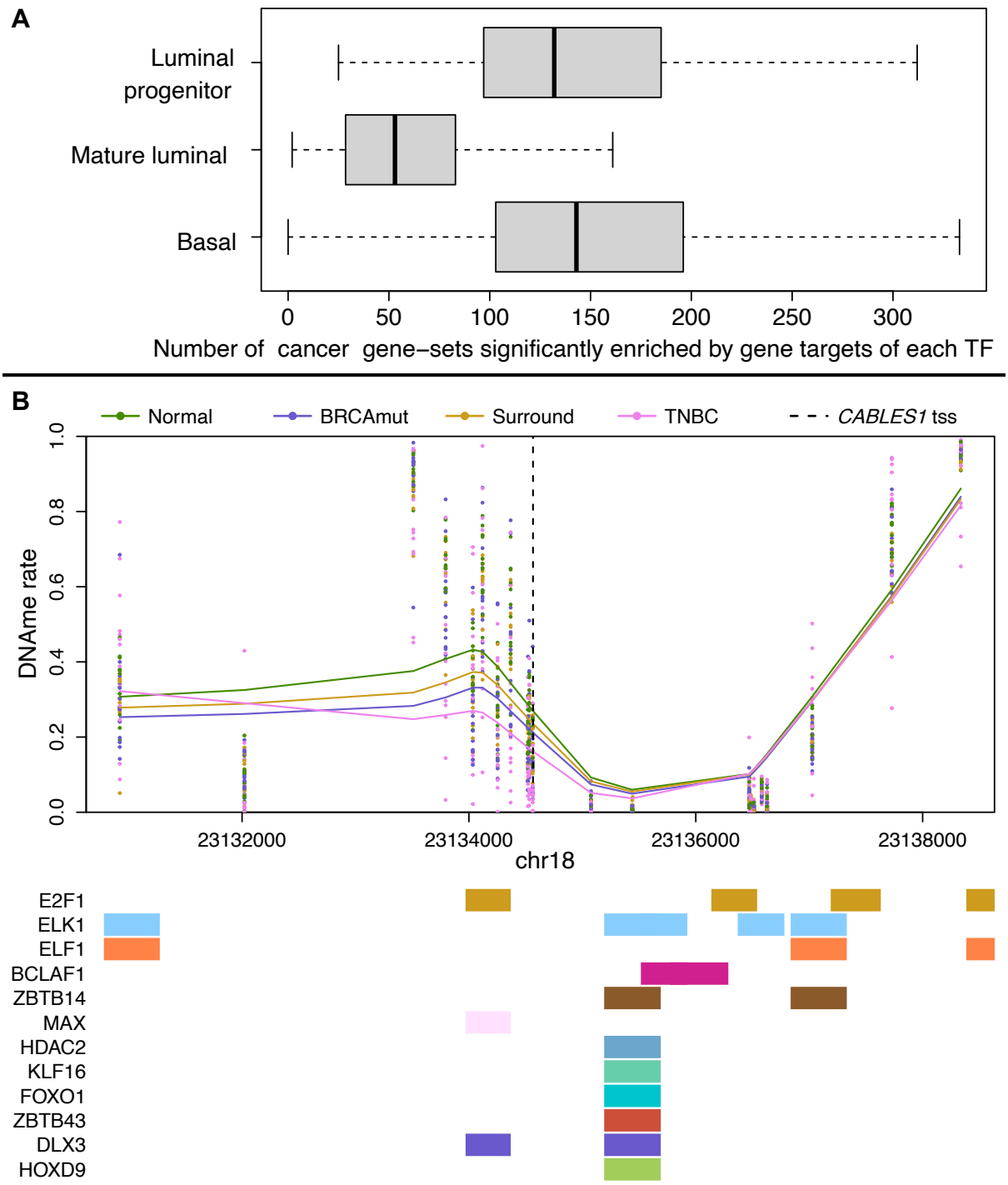

Figure 7: *(a) Boxplots showing numbers of MSigDB cancer gene-sets [49] having significant overlaps with the gene targets of each differentially regulating TF (FDR $p\text{-}val < 0.05$, Fisher's exact test). (b) The cis-regulatory region ($\pm$5kb) around the transcriptional start site (tss) of the CABLES1 gene is shown, highlighting potential cis-regulatory regions for the most important differentially-regulating TFs inferred in luminal progenitor cells as in Sections 2.5, 4.1, and 4.2, and as shown in Figure 6a and b and Table S3; DNAme rates are also shown for previously-published [27] bulk-tissue microarray data from patient volunteers as follows: 14 healthy BRCA wild-type volunteers, 14 healthy BRCA1/2 (FANCS/FANCD2) mutation carriers (BRCAmut), 14 triple-negative breast cancer (TNBC) samples, and 14 samples from the healthy surrounding tissue of TNBC (surr); LOWESS-regression non-linear lines-of-best-fit are also shown.*

carriers compared to wild-type in luminal progenitor cells (Figure 6a and b and Table S3) could show pre-cancerous effects that are present in the methylomes of an at-risk population such as this population of *BRCA* mutation carriers. The gene found to be co-regulated by the largest number of diff-reg TFs in *BRCA* mutation carriers in luminal progenitor cells was *CABLES1*; other examples can be found similarly. Figure 7b shows previously-published [27] bulk-tissue DNAme profiles from healthy, at-risk, and tumour samples over the cis-regulatory region ($\pm$5kb) around the transcriptional start site (tss) of the *CABLES1* gene, along with the potential cis-regulatory regions for the most important diff-reg TFs inferred in luminal progenitor cells. The region of interest can be seen around 23134kb, where there is highly variable methylation level as well as hypomethylation in the TNBC samples (relative to healthy normal tissue). Importantly, some hypomethylation is also visible in the at-risk samples, i.e., in healthy tissue surrounding TNBC, and in healthy tissue from *BRCA1/2* (*FANCS*/*FANCD1*) mutation carriers. To rule out this effect being due to variation in cell-type composition of the samples, we note that there is no significant difference in luminal progenitor proportion in these data comparing healthy normal *BRCA1/2* wild-type and mutation carriers, ($p = 0.182$, *t*-test). Hence, Figure 7b illustrates how corruption to the DNA methylation architecture in precedes malignant transformation in these cells, providing an 'epigenetically permissible environment', also known as an 'epigenetic field defect' [24, 26], or 'epigenetic field cancerisation' [27]. This finding also shows how the GRN inference methodology proposed here could be used to find DNA methylation biomarkers of cancer early-warning. The product of the *CABLES1* gene is a cyclin-dependent kinase and tumour-suppressor that is known to interact with (the best known tumour-supressor) p63 [51], and has previously been shown to be up-regulated in other hormone-driven cancers such as ovarian [52] and endometrial cancers [51]. This proposed GRN inference methodology has also identified the TF-encoding genes (tumour-suppressor regulator) *E2F1*, (cellular proliferation and apoptosis regulator) *MAX*, and (developmental regulator) *DLX3*, as potential master-regulators of the cis-regulatory element around 23134kb where this hypomethylation is visible in cancer and at-risk tissues. Hence, these results show how the GRN inference methodology proposed here can provide functional insights into oncogenic processes for further experimental investigation, for example with targeted molecular interventions.

3 Discussion

In this work, we have shown that the accuracy of inference of GRN TF-gene cis-regulations is improved when prior epigenomic network structure is used in combination with advanced regression methods and gene-expression data. Furthermore, we have shown that DNAme data can be used successfully and effectively to estimate prior epigenomic network structure. There is a wealth of existing and publicly-available DNAme datasets for a wide range of tissue types that can now be leveraged using our proposed methodology to improve inference of gene regulatory processes [13]. This contrasts with novel ATAC-seq datasets that are expensive to generate, and are currently less numerous in the public domain. However, as the availability increases of single-cell ATAC-seq data and of matched single-cell ATAC-seq and RNA-seq data, it will be interesting to see how our proposed methodology may be complemented and enhanced by incorporation of single-cell chromatin accessibility data. For example, we note that single-cell ATAC-seq data are typically very sparse. Therefore, a large number of cis-regulatory regions would be expected to have low coverage in such data, and particularly in these regions the DNAme epigenomic prior network may provide useful complementary information.

The inferences made when our proposed methodology is applied to real data are in line with established and well-validated biological models. We observed that DNA loci in cis-regulatory regions of interest typically show highly variable methylation levels, in line with previous observations in DNA methylation data [28, 29], in single-cell RNA-seq data [53–56], and in proteomic data [57]. We also found that differential gene regulation at the epiblast - hypoblast (primitive endoderm, PE) lineage bifurcation is characterised by an upregulation of transcription factors in PE relative to epiblast (Figure 5), reflecting a cell fate choice in PE, whilst epiblast (Epi) remains able to choose from multiple lineage paths [43]. A similar, but more pronounced pattern in *BRCA1* (*FANCS*) mutation carriers compared to the whole population is also seen in luminal progenitor and basal cells (Figure 6a, b, e, and f). This can be interpreted as evidence that these cells in *BRCA1* mutation carriers retain more ability to make multiple fate choices, for example acquiring characteristics of aggressive cancer, when compared with these cell types in the general population. We note in particular that luminal progenitor cells are thought to be the cells of origin of highly aggressive triple-negative breast cancer (TNBC) [27, 45].

Further investigation of the differentially regulating transcription factors and their targets shown in

Figure 6a, b, e, and f may allow gene regulatory processes to be uncovered that are crucial to the oncogenic potential of luminal progenitor and basal cells in *BRCA1* (*FANCS*) and *BRCA2* (*FANCD1*) mutation carriers, as well giving clues about possible pharmacological interventions. For example, Figure 7b shows how the methodology proposed here can be used to identify CpG loci that show altered methylation levels prior to cancer formation in at-risk tissues and populations. Hence, testing the DNAme rate of CpGs in at-risk individuals in certain cis-regulatory regions of tumour suppressor genes, identified by the methodology proposed here, could provide important early-warning biomarkers of cancer or of cancer predisposition. However, further work will be required to systematically find and then validate all the examples of changes in DNAme patterns that may be discovered with our proposed methodology, and that are associated with changes in TF cis-regulation of specific genes that precede the onset of disease.

4 Methods

4.1 Epigenomic prior network construction

To obtain the potential TF cis-regulators of a particular gene, we build on established methods called lever [14] and cistarget [15]. Those methods examine whether the closeness of matching between TF binding motifs and cis-regulatory DNA sequences corresponds to sample-specific observations at the same cis-regulatory DNA regions, such as open chromatin (as inferred from ATAC-seq or LiCAT-seq data). Lever [14] and cistarget [15] approach this as a prediction problem: i.e., does the closeness of TF motif match to the DNA sequence at a particular location predict the epigenomic patterns (such as open chromatin) that are observed at that location. In lever [14] and cistarget [15], the accuracy of this prediction is measured by the AUC (area under receiver-operator curve) statistic, which is related to Somers' D statistic by the transformation $D = 2 \times \text{AUC} - 1$ [58]. If the correlation coefficient $\rho \in [-1, 1]$ is a measure of association between two binary variables, then $D \in [-1, 1]$ can be thought of as an analogous measure of association between a binary variable and a continuous variable. From the perspective of the AUC statistic, the binary variable corresponds to a representation of the true class $y^{(r)} \in \{0, 1\}$ for rankings $r \in \{1, ..., R\}$, and the continuous variable $x^{(r)}$ is used to predict that class, where a value of the AUC statistic close to 1 indicates that high rankings mostly correspond to a true class of 1.

In MethCorTarget, we use the absolute Spearman correlation coefficient to assess whether the average DNAme rate $y_j^{(r)} \in [0, 1]$ for sample $j \in \{1, ..., n\}$ at cis-regulatory region $r \in \{1, ..., R\}$ corresponds to the closeness of the match $x_l^{(r)}$ between TF motif $l \in \{1, ...L\}$ and the DNA sequence at that region, where correlation coefficient $\rho_{l,j}(x_l^{(r)}, y_j^{(r)})$ is calculated over $r \in \{1, ..., R\}$. In modified lever-cistarget (MLC), we equivalently use Somers' D to test whether the presence of open chromatin (assessed by LiCAT-seq or H3K27ac) at a cis-regulatory region corresponds to the closeness of the match between the TF motif and the DNA sequence at that region. After calculating $\rho_{l,j}$ (for a given DNAme sample j) or $D_{l,j}$ (for other epigenomic data sample j) for TF motif l, we apply the Fisher transform to obtain approximately standard Gaussian-distributed variables $z_{l,j} \sim \mathcal{N}(0, 1)$ [59], $l \in \{1, ...L\}$, $j \in \{1, ..., n\}$, for L TFs and n samples (assuming the distributional assumptions also hold sufficiently well for D), and then calculate corresponding p-values from a standard Gaussian distribution. We then retain TFs in the subsequent analysis if the corresponding z-test leads to p-value $p_{l,j} < 0.1$, noting that this threshold provides optimum identifiability of network structures [60]. We also note that this p-value threshold might result in many false-positives for a large number of TFs, but that many of these would be expected to be filtered out in the next step in the GRN inference using gene expression data with advanced regression methods. The aim of the epigenomic prior network is to identify potential regulators of a target-gene, which are then chosen from by the advanced regression method as it is applied to gene-expression data. We refer to this method of epigenomic prior network inference as MLC (modified-lever-cistarget) when the method calculates Somers' D from chromatin accessibility data (or similar), and we refer to this method as MethCorTarget when the method calculates the Spearman correlation coefficient ρ from DNAme data.

4.2 Gene regulatory network inference

Gene regulatory network (GRN) inference was carried out to infer TF-gene regulations by modelling the expression of each target-gene as a (possibly non-linear) combination of the expression of a set of potential regulators (Equation 1) [4]. The target-gene approach to GRN inference [4] uses advanced regression methods such as random forests regression [61], and recently the xgboost implementation

of random forests as gradient-boosted trees [16], to model the expression level of a particular 'target' gene (the response), conditional on the expression levels of a set of other genes (predictors, such as TFs), Eq.1. The target-gene approach to GRN inference is best known in the context of GENIE3 [5] and GRNboost2 [6]. In GENIE3 and GRNboost2, the set of potential regulators corresponds to all the TFs in the dataset of interest, or to some other pre-specified list of potential regulators that importantly is the same for all target-genes. In our proposed methodology, we use the epigenomic prior network to first infer a specific set of potential TF regulators that can be different for each gene (Figure 1). This improves the accuracy of inference by the advanced regression method (Figure 2), because when using advanced regression such as xgboost for variable selection, the advanced regression only needs to choose from 10s of possible cis-regulating TFs. These 10s of potential cis-regulators correspond to prior epigenomic information about possible interactions of the TF with the cis-regulatory DNA of the target-gene. This contrasts with the full list of 100s or 1000s of TFs that would need to be specified genome-wide when using GENIE3 [5] and GRNboost2 [6], as part of the SCENIC / SCENIC+ pipeline [1–3]. Finally, we construct the GRN by combining the local target-gene network model fits genome-wide, with each gene taking a turn as the response variable in the advanced regression, i.e., as the target-gene [4].

We found previously [12] that advanced regression methods as part of the target-gene approach to GRN inference perform best with normalisation methods such as FPKM and TPM. In this work, we chose to use CPM normalisation rather than FPKM or TPM normalisation, because inferences from advanced regression with the random forests method should not be affected by monotonic transformations of the data, such as scaling by gene length. In comparison with that previous work [12], we also chose to use random forests regression [61] (xgboost implementation [16]) rather than mutual information, for three reasons as follows: (1) The computational efficiency of xgboost [16]) makes it feasible to use xgboost with large, modern single-cell datasets such as the breast cancer at-risk dataset we analyse in this work. This contrasts with the much smaller number of cells in the human embryonic development dataset, where cell numbers are restricted due to the very precious and restricted nature of the source material; (2) In our previous work [12] we used the GENIE3 implementation [5] of target-gene GRN inference with random-forests [61], but in this work we have have shown a clear improvement over GENIE3 on synthetic data by using prior network structure (Figure 2), noting that GENIE3 does not use prior network structure that is tailored to each target-gene; (3) In this work we have based our prior epigenomic network inference methods on the lever [14] / cistarget [15] approach, which is also used together with random forests [61] as part of the well validated scenic [1, 3] pipeline. On the other hand our previous work [12] used RGT-HINT [35] to infer potential TF-gene cis-regulations.

4.3 Dataset details and pre-processing

LiCAT-seq data processing

For the human embryonic development LiCAT-seq dataset [37], the DNA loci of cis-regulatory TF motifs corresponding to open chromatin were inferred as part of an earlier study [12] (with alignment to the GRCh38 reference human genome), and were reused in this work. These TF motif footprints at open chromatin were originally inferred from the human embryonic development LiCAT-seq dataset available from the NCBI Sequence Read Archive (SRA) under accession number SRP163205, from the URL `https://www.ncbi.nlm.nih.gov/sra/?term=SRP163205`
The TF-motif footprints at open chromatin were originally obtained using rgt-hint v0.13.0 [35] with TF binding models from HOCOMOCO and JASPAR [37], finding 433270 potential cis-regulations identified over 13950 different genes by 894 different TFs.

We applied pycistarget and modified-lever-cistarget (MLC) to the downstream peak-calls (54425 peak-calls genome wide) for the human embryonic development LiCAT-seq dataset [37] for comparison. When applied to these peak-calls, pycistarget returned 124 significant TF motifs across 39298 distinct genomic regions. The MLC method found 549 significant TF motifs, matched across 27280 distinct genomic regions (z-test, $p < 0.1$, we note that many false positives generated in this step are expected to be removed by the subsequent GRN inference with gene-expression data). For MLC, potential cis-regulations at these 27280 genomic regions were inferred from the closest gene promoters (< 5kb, according to Ensemble 107 [62]). This lead to a prior epigenomic network for the human embryonic development LiCAT-seq dataset, representing 21381 possible cis-regulations of 2650 different genes by 270 different TFs.

H3K27ac ChIP-seq data processing

For the breast cancer at-risk H3K27ac ChIP-seq dataset [46], raw fastq files of unaligned sequencer reads for purified cell samples for the cell types luminal progenitor, mature luminal, and basal from each of 3 women (aged 22, 37 and 41), were downloaded from EGA (the European Genome-phenome Archive, `https://ega-archive.org`), under accession number EGAS00001000552. ChIP-seq reads were aligned to the GRCh38 (hg38) reference human genome assembly from Ensembl 109 [63] using the Burrows-Wheeler Alignment (BWA) software [64] (with default settings). Significant ChIP-seq peaks (FDR-adj $p < 0.05$) were then called using the callpeak function from the macs2 software [65, 66] (with default settings), resulting in three sets of peak calls per cell type. Applying MLC to these data for the cell types luminal progenitor, mature luminal, and basal, lead to (respectively) 583, 591, and 595 significant TF motifs matched across 25706, 27668, and 22570 genomic regions (z-test, $p < 0.1$, we note that many false positives generated in this step are expected to be removed by the subsequent GRN inference with gene-expression data). For each genomic region, potential cis-regulations were inferred from the closest gene promoters (< 5kb, according to Ensemble 107 [62]), also requiring that the potential cis-regulation should be observed at this significance level in a majority of the samples for that cell type. This lead to a prior epigenomic network for each of the cell types luminal progenitor, mature luminal, and basal, representing (respectively) 142766, 142237, and 110513 possible cis-regulations of 4222, 4194, and 3597 different genes, by 282, 285, and 277 different TFs.

DNAme bisulphite-seq data processing

For the human embryo DNAme dataset [38], CpG read-count data tables summarising bisulphite-seq libraries (aligned to the hg19 reference human genome) for 103 human ICM cell samples (inner cell mass, comprising epiblast and primitive endoderm cells, one library per cell sample), were downloaded from from the Gene Expression Omnibus (GEO) under accession number GSE81233, from the URL: `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81233`

For the breast cancer at-risk DNAme dataset [46], bisulphite-seq DNAme data for purified cell samples donated by 3 women (aged 22, 37 and 41, one library per cell type for each woman) for luminal progenitor, mature luminal, and basal cells, were downloaded from EGA (the European Genome-phenome Archive, `https://ega-archive.org`), under accession number EGAS00001000552. The downloaded raw sequencer reads were aligned to the GRCh37 (hg19) reference human genome assembly from Ensembl 104 [67], and methylated and unmethylated reads were counted at CpG sites, using the Bismark software [68] (with default settings).

To generate the epigenomic prior networks from these DNAme datasets, MethCorTarget was applied to the raw methylated and unmethylated bisulphite-seq counts, as follows. First, the methylated and unmethylated counts were aggregated over all candidate TF-motif binding regions that were mapped by at least 10 reads in total, leading to an average DNAme rate for each candidate TF-motif binding region, for each sample (i.e., bisulphite-seq library). The average methylation rates over the samples were then calculated for all TF-motif binding regions, for only those regions with data for at least 3 cells in the human embryonic development dataset, and for only those regions with data for the majority of samples in the breast cancer at-risk dataset. We note that DNAme data are typically quite noisy, but by averaging the methylation rates over TF binding regions, we mitigate some of the effect of this noise.

We used MethCorTarget to test the correlation of the ranking of the TF-motif DNA sequence matches for each binding region, with the average methylation rates over the binding regions calculated as above. We applied MethCorTarget in the same way for ICM cells of the human embryonic development dataset, and for the cell types luminal progenitor, mature luminal, and basal cells of the breast cancer at-risk dataset. For ICM cells, MethCorTarget found 868 significant TF motifs matched across 95122 genomic regions. For luminal progenitor, mature luminal, and basal cells, MethCorTarget found (respectively), 826, 780, and 747 significant TF motifs matched across 206076, 201359, and 195211 genomic regions (z-test, $p < 0.1$; we note that many false positives generated in this step are expected to be removed by the subsequent GRN inference with gene-expression data). For each genomic region, potential cis-regulations were inferred from the closest gene promoters (< 5kb) to the motif-matched TFs (according to Ensemble 107 [62]). This lead to a prior epigenomic network for ICM representing 417320 possible cis-regulations of 23575 genes by 375 TFs. It also lead to a prior epigenomic network for each of the cell types luminal progenitor, mature luminal, and basal, representing (respectively) 1035073, 946625, and 920222 possible cis-regulations of 34792, 34680, and 34021 genes, by 413, 370, and 382 TFs.

For plotting the bisulphite-seq reads aligned to hg19 on GRCh38 (hg38) co-ordinates, the liftover

package in R was used to convert CpG loci from the hg19 coordinates to hg38.

RNA-seq data processing

The human embryonic development RNA-seq dataset [40] was downloaded from the ArrayExpress repository under accession number E-MTAB-3929, from URL:
`https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-3929`
Raw counts of 26178 RNA transcripts in 1529 cells were downloaded, corresponding to cells from the blastocyst stage of embryonic development. After quality control and filtering to retain cells from embryos at 5-7 days post-fertilisation, data for 20407 features (RNA transcripts) in 1258 samples (cells) remained.

The breast cancer at-risk RNA-seq dataset [50] was downloaded from the Gene Expression Omnibus (GEO) repository under accession number GSE161529, from URL:
`https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161529`
Raw counts of 33538 RNA transcripts in 26272 cells were downloaded, corresponding to cells from the epithelial lineage and supporting cells from pre-menopausal and nulliparous women, of whom $n = 2$ are *BRCA1* (*FANCS*) mutation carriers (BRCAmut) and $n = 5$ are *BRCA1* wild-type (WT). After quality control, data for 14835 features (RNA transcripts) in 17730 samples (cells) remained, and were used for the subsequent GRN inference.

Quality control was carried out as follows, for both the human embryonic development dataset and the breast cancer at-risk dataset. High quality features (RNA transcripts) and samples (cells) were retained by filtering out any RNA transcripts with non-zero counts in fewer than 50 cells, and by filtering out any cells with non-zero counts for fewer than 750 RNA transcripts, filtering out cells for which the most highly expressed transcript excluding MALAT1 account for at least 10% of the total counts, filtering out cells for which at least 10% of the total counts correspond to transcripts encoding mitochondrial proteins, and filtering out cells for which at least 50% of the total counts correspond to transcripts encoding ribosomal proteins. For the human embryo dataset, features were instead filtered out if they have non-zero counts in fewer than 10 samples (cells), noting the smaller sample size for this dataset; also for he human embryo dataset no filtering was carried out based on transcripts encoding mitochondrial proteins because they did not appear in the dataset. For the breast cancer at-risk dataset, batch correction using the COMBAT software [69] was also applied to adjust for batch effects.

After quality control, differential expression analysis was carried out using the limma software [70] from the edgeR package [71] in R, with significance level false discovery rate-adjusted $p < 0.05$. Separately after quality control, the top features were selected for dimension reduction, visualisation, and clustering, according to an estimator of the amount of biological variability compared to technical variability [72] in each feature $i \in \{1, ..., p\}$ defined as $\log V_i / \log m_i$, where V_i and m_i are the empirical variance and mean for feature i, respectively. The top 2000 features were retained in this way for the breast cancer at-risk dataset, and the top 5000 features were retained for the embryonic development dataset (recognising that more features are needed for estimating the Laplacian eigenspace as this dataset contains many fewer cell samples). We note that this procedure in practice approximates a standard technique for obtaining the features with the greatest biological variability in which a line of best fit is found for a plot of $\log V_i$ against $\log m_i$, and the features that are furthest above this line are then selected. This practice is based on the assumption that the data for feature i will follow a Poisson distribution when only technical variation is present meaning that $V_i \approx m_i$, and will be overdispersed when biological variation is present meaning that $V_i \approx m_i + \phi_i m_i^2$, where $\phi_i > 0$ is an overdispersion parameter for feature i [72]. We note that there is currently no consensus on the best-practice for selecting the most informative features using models of overdispersion, and that the best solution is likely to depend both on the downstream processing and on as the datasets that are analysed [53–56].

The Laplacian eigenspace (LE) is the basis for the visualisations and clustering shown in Figure S1, Figure S2, and Figure S4, and is estimated as the subspace of the data defined by the eigendecomposition of the graph-Laplacian representation of the datapoints [73], according to the eigenvectors that correspond to the eigenvalues that explain at least 1% of the total variance of the data, after removing the mean component. The single-cell data for the specific cell types analysed in the GRN inference are identified from the LE with unsupervised learning by fitting a Gaussian mixture model in the Laplacian eigenspace (GMM-LE clustering) [73, 74], as shown in UMAP projections [75] from the LE [73] that appear in Figure S1, Figure S2, and Figure S4.

For the human embryonic development dataset, the epiblast cells used in the GRN inference are

the same ones that we used in previous work [73], identified as cluster 5 in Figure S1; the identifiers for these Epi cells from the original study that generated those data [40] are listed in Table S6. GMM-LE clustering was then applied hierarchically to cluster 2 from Figure S1, resulting in 37 primitive endoderm (PE) cells identified as cluster 4 in Figure S2; the identifiers for these PE cells from the original study that generated these data [40] are listed in Table S7. The 68 and 37 Epi and PE cells (respectively) are validated in Figure S3 by the expression levels of known Epi marker genes *NANOG* and *KLF17* and PE marker genes *GATA4* and *SOX17* [12], and are used as representative samples of gene-expression profiles of those cell types in the GRN inference. For the breast cancer at-risk dataset, we previously identified 5531 luminal progenitor, 2535 mature luminal, and 3913 basal cells [73], which are identified as clusters 1, 2, and 8, respectively in Figure S4. These cell samples are then used as representative examples of gene-expression profiles of those cell types in the GRN inference.

DNAme bulk microarray data processing

Bulk DNAme data were analysed from samples from volunteers as follows: $n = 14$ healthy *BRCA* wild-type, $n = 14$ healthy *BRCA1/2* mutation carriers, $n = 14$ triple-negative breast cancer (TNBC) samples, and $n = 14$ samples from the healthy surrounding tissue of TNBC (as originally described in [27]); these data were preprocessed as follows. Raw DNAme idat data files were loaded in R and background-corrected and normalised with BMIQ [76]. Probes were removed if they had $< 95\%$ coverage across samples, and any remaining probes with detection p-value > 0.05 were replaced by k-NN imputation, with $k = 5$.

TF-motif databases

The rankings for the matches of TF-motif to DNA sequence location for human, that we used with pycistarget, MLC (modified-lever-cistarget) and MethCorTarget, are based on 32765 unique TF motifs that were originally collated from 29 different motif databases as part of the SCENIC+ package [3], and were downloaded under the filenames
`hg38_screen_v10_clust.regions_vs_motifs.rankings.feather` (for the motif rankings) and
`motifs-v10nr_clust-nr.mgi-m0.001-o0.0.tbl` (for the TF-motif matches), from the URL:
`https://resources.aertslab.org/cistarget/`

4.4 Synthetic data generation

Synthetic gene expression data was generated, corresponding to known ground-truth gene regulatory network structure. These data were generated using the 'gene netweaver' software [77], which uses realistic differential-equations models to simulate the data according to models of chemical reaction kinetics; this software was used previously to generate data for the DREAM4 and DREAM5 network inference challenges [36] that were originally won by GENIE3 [5]. Using gene-net weaver, we generated 300 synthetic gene-expression profiles based on known gene-regulatory network structure. These 300 synthetic gene-expression profiles were then analysed equivalently to 300 normalised single-cell libraries sequenced at high depth. The number of synthetic data samples, the maximum synthetic network size, and the number of potential regulators of each target gene, were determined by the constraints of the data generating process

4.5 Gene-set enrichment analysis

The cancer gene-sets were downloaded from the Broad Institute's Molecular Signatures Database (MSigDB) [49] in January 2024. The epiblast gene-sets were validated and used extensively in previously published work [12].

4.6 Deconvolution of bulk DNA methylation data

Relative proportions of cell-types fibroblast, immune and fat representing important tissue constituents as well as the epithelial subtypes luminal progenitor, mature luminal, and basal cells, were estimated in the bulk DNAme data using the 'Breast Epithelial Subtypes' package in R [34].

5 References

[1] Aibar S, González-Blas CB, Moerman T, Huynh-Thu VA, Imrichova H, Hulselmans G, et al. SCENIC: single-cell regulatory network inference and clustering. Nature methods. 2017;14(11):1083-6.

[2] Van de Sande B, Flerin C, Davie K, De Waegeneer M, Hulselmans G, Aibar S, et al. A scalable SCENIC workflow for single-cell gene regulatory network analysis. Nature Protocols. 2020;15(7):2247-76.

[3] Bravo González-Blas C, De Winter S, Hulselmans G, Hecker N, Matetovici I, Christiaens V, et al. SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. Nature Methods. 2023;20(9):1355-67.

[4] Dobra A, Hans C, Jones B, Nevins JR, Yao G, West M. Sparse graphical models for exploring gene expression data. Journal of Multivariate Analysis. 2004;90(1):196-212.

[5] Huynh-Thu VA, Irrthum A, Wehenkel L, Geurts P. Inferring regulatory networks from expression data using tree-based methods. PloS one. 2010;5(9):e12776.

[6] Moerman T, Aibar Santos S, Bravo González-Blas C, Simm J, Moreau Y, Aerts J, et al. GRNBoost2 and Arboreto: efficient and scalable inference of gene regulatory networks. Bioinformatics. 2019;35(12):2159-61.

[7] Brandman O, Meyer T. Feedback loops shape cellular signals in space and time. Science. 2008;322(5900):390-5.

[8] Bartlett TE, Olhede SC, Zaikin A. A DNA methylation network interaction measure, and detection of network oncomarkers. PloS one. 2014;9(1):e84573.

[9] Bartlett TE, Zaikin A, et al. Detection of epigenomic network community oncomarkers. The Annals of Applied Statistics. 2016;10(3):1373-96.

[10] Bartlett TE, Kosmidis I, Silva R. Two-way sparsity for time-varying networks with applications in genomics. The Annals of Applied Statistics. 2021;15(2):856-79.

[11] Bartlett T. Fusion of single-cell transcriptome and DNA-binding data, for genomic network inference in cortical development. BMC bioinformatics. 2021;22(1):1-19.

[12] Alanis-Lobato G, Bartlett TE, Huang Q, Simon C, McCarthy A, Elder K, et al. MICA: A multi-omics method to predict gene regulatory networks in early human embryos. Life Science Alliance. 2023;7(1):e202302415.

[13] Loyfer N, Magenheim J, Peretz A, Cann G, Bredno J, Klochendler A, et al. A DNA methylation atlas of normal human cell types. Nature. 2023;613(7943):355-64.

[14] Warner JB, Philippakis AA, Jaeger SA, He FS, Lin J, Bulyk ML. Systematic identification of mammalian regulatory motifs' target genes and functions. Nature methods. 2008;5(4):347-53.

[15] Herrmann C, Van de Sande B, Potier D, Aerts S. i-cisTarget: an i ntegrative genomics method for the prediction of regulatory features and cis-regulatory modules. Nucleic acids research. 2012;40(15):e114-4.

[16] Chen T, Guestrin C. Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining; 2016. p. 785-94.

[17] Butte AJ, Kohane IS. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Biocomputing 2000. World Scientific; 1999. p. 418-29.

[18] Daub CO, Steuer R, Selbig J, Kloska S. Estimating mutual information using B-spline functions–an improved similarity measure for analysing gene expression data. BMC bioinformatics. 2004;5(1):1-12.

[19] Margolin AA, Nemenman I, Basso K, Wiggins C, Stolovitzky G, Dalla Favera R, et al. ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. BMC bioinformatics. 2006;7(1):1-15.

[20] Faith JJ, Hayete B, Thaden JT, Mogno I, Wierzbowski J, Cottarel G, et al. Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles. PLoS biology. 2007;5(1):e8.

[21] Chan TE, Stumpf MP, Babtie AC. Gene regulatory network inference from single-cell data using multivariate information measures. Cell systems. 2017;5(3):251-67.

[22] Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. science. 2003;302(5643):249-55.

[23] Bartlett TE, Müller S, Diaz A. Single-cell co-expression subnetwork analysis. Scientific reports. 2017;7(1):15066.

[24] Bernstein C, Nfonsam V, Prasad AR, Bernstein H. Epigenetic field defects in progression to cancer. World journal of gastrointestinal oncology. 2013;5(3):43.

[25] Bartlett TE, Chindera K, McDermott J, Breeze CE, Cooke WR, Jones A, et al. Epigenetic reprogramming of fallopian tube fimbriae in BRCA mutation carriers defines early ovarian cancer evolution. Nature communications. 2016;7(1):1-10.

[26] Teschendorff AE, Gao Y, Jones A, Ruebner M, Beckmann MW, Wachter DL, et al. DNA methylation outliers in normal breast tissue identify field defects that are enriched in cancer. Nature communications. 2016;7(1):1-12.

[27] Bartlett TE, Evans I, Jones A, Barrett JE, Haran S, Reisel D, et al. Antiprogestins reduce epigenetic field cancerization in breast tissue of young healthy women. Genome medicine. 2022;14(1):1-18.

[28] Bartlett TE, Zaikin A, Olhede SC, West J, Teschendorff AE, Widschwendter M. Corruption of the intra-gene DNA methylation architecture is a hallmark of cancer. PloS one. 2013;8(7):e68285.

[29] Bartlett TE, Jones A, Goode EL, Fridley BL, Cunningham JM, Berns EM, et al. Intra-gene DNA methylation variability is a clinically independent prognostic marker in women's cancers. PloS one. 2015;10(12):e0143178.

[30] Gosden RG, Feinberg AP. Genetics and epigenetics—nature's pen-and-pencil set. Mass Medical Soc; 2007.

[31] Kaluscha S, Domcke S, Wirbelauer C, Stadler MB, Durdu S, Burger L, et al. Evidence that direct inhibition of transcription factor binding is the prevailing mode of gene and repeat repression by DNA methylation. Nature genetics. 2022;54(12):1895-906.

[32] Feinberg AP, Ohlsson R, Henikoff S. The epigenetic progenitor origin of human cancer. Nature reviews genetics. 2006;7(1):21-33.

[33] Teschendorff AE, Breeze CE, Zheng SC, Beck S. A comparison of reference-based algorithms for correcting cell-type heterogeneity in Epigenome-Wide Association Studies. BMC bioinformatics. 2017;18(1):105.

[34] Bartlett TE, Jia P, Chandna S, Roy S. Inference of tissue relative proportions of the breast epithelial cell types luminal progenitor, basal, and luminal mature. Scientific reports. 2021;11(1):1-10.

[35] Li Z, Schulz MH, Look T, Begemann M, Zenke M, Costa IG. Identification of transcription factor binding sites using ATAC-seq. Genome biology. 2019;20(1):1-21.

[36] Meyer P, Saez-Rodriguez J. Advances in systems biology modeling: 10 years of crowdsourcing DREAM challenges. Cell Systems. 2021;12(6):636-53.

[37] Liu L, Leng L, Liu C, Lu C, Yuan Y, Wu L, et al. An integrated chromatin accessibility and transcriptome landscape of human pre-implantation embryos. Nature communications. 2019;10(1):364.

[38] Zhu P, Guo H, Ren Y, Hou Y, Dong J, Li R, et al. Single-cell DNA methylome sequencing of human preimplantation embryos. Nature genetics. 2018;50(1):12-9.

[39] Blakeley P, Fogarty NM, Del Valle I, Wamaitha SE, Hu TX, Elder K, et al. Defining the three cell lineages of the human blastocyst by single-cell RNA-seq. Development. 2015;142(18):3151-65.

[40] Petropoulos S, Edsgärd D, Reinius B, Deng Q, Panula SP, Codeluppi S, et al. Single-cell RNA-seq reveals lineage and X chromosome dynamics in human preimplantation embryos. Cell. 2016;165(4):1012-26.

[41] Allègre N, Chauveau S, Dennis C, Renaud Y, Meistermann D, Estrella LV, et al. NANOG initiates epiblast fate through the coordination of pluripotency genes expression. Nature Communications. 2022;13(1):3550.

[42] Ben-Porath I, Thomson MW, Carey VJ, Ge R, Bell GW, Regev A, et al. An embryonic stem cell–like gene expression signature in poorly differentiated aggressive human tumors. Nature Genetics. 2008;40(5):499-507.

[43] Guo G, Stirparo GG, Strawbridge SE, Spindlow D, Yang J, Clarke J, et al. Human naive epiblast cells possess unrestricted lineage potential. Cell stem cell. 2021;28(6):1040-56.

[44] Kuchenbaecker KB, McGuffog L, Barrowdale D, Lee A, Soucy P, Dennis J, et al. Evaluation of polygenic risk scores for breast and ovarian cancer risk prediction in BRCA1 and BRCA2 mutation carriers. JNCI: Journal of the National Cancer Institute. 2017;109(7):djw302.

[45] Tharmapalan P, Mahendralingam M, Berman HK, Khokha R. Mammary stem cells and progenitors: targeting the roots of breast cancer for prevention. The EMBO journal. 2019;38(14):e100852.

[46] Pellacani D, Bilenky M, Kannan N, Heravi-Moussavi A, Knapp DJ, Gakkhar S, et al. Analysis of normal human mammary epigenomes reveals cell-specific active enhancer states and associated transcription factor networks. Cell reports. 2016;17(8):2060-74.

[47] Creyghton MP, Cheng AW, Welstead GG, Kooistra T, Carey BW, Steine EJ, et al. Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proceedings of the National Academy of Sciences. 2010;107(50):21931-6.

[48] Bogdanović O, Fernandez-Miñán A, Tena JJ, de la Calle-Mustienes E, Hidalgo C, van Kruysbergen I, et al. Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome research. 2012;22(10):2043-53.

[49] Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences. 2005;102(43):15545-50.

[50] Pal B, Chen Y, Vaillant F, Capaldo BD, Joyce R, Song X, et al. A single-cell RNA expression atlas of normal, preneoplastic and tumorigenic states in the human breast. The EMBO journal. 2021;40(11):e107333.

[51] Huang JR, Tan GM, Li Y, Shi Z. The emerging role of cables1 in cancer and other diseases. Molecular pharmacology. 2017;92(3):240-5.

[52] Morin PJ. Learning about ovarian tumorigenesis by watching Cables. Cancer biology & therapy. 2008;7(2):189-90.

[53] Townes FW, Hicks SC, Aryee MJ, Irizarry RA. Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. Genome biology. 2019;20(1):1-16.

[54] Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome biology. 2019;20(1):296.

[55] Lause J, Berens P, Kobak D. Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. Genome biology. 2021;22(1):22.

[56] Choudhary S, Satija R. Comparison and evaluation of statistical error models for scRNA-seq. Genome biology. 2022;23(1):27.

[57] Mrukwa G, Polanska J. DiviK: divisive intelligent K-means for hands-free unsupervised clustering in big biological data. BMC bioinformatics. 2022;23(1):538.

[58] Newson R. Parameters behind “nonparametric” statistics: Kendall’s tau, Somers’ D and median differences. The Stata Journal. 2002;2(1):45-64.

[59] Fisher RA. Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. Biometrika. 1915:507-21.

[60] Gallagher I, Jones A, Bertiger A, Priebe CE, Rubin-Delanchy P. Spectral embedding of weighted graphs. Journal of the American Statistical Association. 2023:1-10.

[61] Breiman L. Random forests. Machine learning. 2001;45:5-32.

[62] Cunningham F, Allen JE, Allen J, Alvarez-Jarreta J, Amode MR, Armean IM, et al. Ensembl 2022. Nucleic acids research. 2022;50(D1):D988-95.

[63] Martin FJ, Amode MR, Aneja A, Austine-Orimoloye O, Azov AG, Barnes I, et al. Ensembl 2023. Nucleic acids research. 2023;51(D1):D933-41.

[64] Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. bioinformatics. 2009;25(14):1754-60.

[65] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome biology. 2008;9(9):1-9.

[66] Gaspar JM. Improved peak-calling with MACS2. BioRxiv. 2018:496521.

[67] Howe KL, Achuthan P, Allen J, Allen J, Alvarez-Jarreta J, Amode MR, et al. Ensembl 2021. Nucleic acids research. 2021;49(D1):D884-91.

[68] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. bioinformatics. 2011;27(11):1571-2.

[69] Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118-27.

[70] Smyth GK. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology. 2004;3(1).

[71] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. bioinformatics. 2010;26(1):139-40.

[72] McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic acids research. 2012;40(10):4288-97.

[73] Bartlett TE, Chandna S, Roy S. Stochastic networks theory to model single-cell genomic count data. arXiv preprint arXiv:230302498. 2023.

[74] Rubin-Delanchy P, Cape J, Tang M, Priebe CE, et al. A statistical interpretation of spectral embedding: the generalised random dot product graph. Journal of the Royal Statistical Society Series B. 2022;84(4):1446-73.

[75] McInnes L, Healy J, Melville J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:180203426. 2018.

[76] Teschendorff AE, Marabita F, Lechner M, Bartlett T, Tegner J, Gomez-Cabrero D, et al. A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. Bioinformatics. 2013;29(2):189-96.

[77] Schaffter T, Marbach D, Floreano D. GeneNetWeaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics. 2011;27(16):2263-70.

6 Additional information

Availability of data and materials

The human embryonic development DNAme dataset is available from the Gene Expression Omnibus (GEO) under accession number GSE81233, from the URL `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE81233`

The human embryonic development LiCAT-seq dataset is available from the NCBI Sequence Read Archive (SRA) under accession number SRP163205, from the URL `https://www.ncbi.nlm.nih.gov/sra/?term=SRP163205`

The breast cancer at-risk DNAme (bisulphite-seq), and H3K27ac (ChIP-seq) datasets, are available from EGA under accession number EGAS00001000552, from the URL `https://ega-archive.org`

The human embryonic development single-cell RNA-seq dataset is available from ArrayExpress under accession number E-MTAB-3929, from the URL `https://www.ebi.ac.uk/biostudies/arrayexpress/studies/E-MTAB-3929`

The breast cancer at-risk single-cell RNA-seq dataset is available from the Gene Expression Omnibus (GEO) repository under accession number GSE161529, from the URL `https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161529`

The breast cancer at-risk bulk DNAme dataset and associated demographic information is available from EGA under accession number EGAS00001005070, from the URL `https://egaarchive.org/studies/EGAS00001005070`

The TF-motif DNA-sequence match rankings are available from the URL: `https://resources.aertslab.org/cistarget/`

Software

Downstream analyses were carried out in R (versions 4.1-4.3), except for the application of pycistarget to the LiCAT-seq data, which used Python version 3.10 and pycistarget version 1.0. The 'Breast Epithelial Subtypes' R package is available from
https://github.com/tombartlett/BreastEpithelialSubtypes

R scripts implementing the epigenomic prior network construction, gene regulatory network inference, and differential cis-regulatory analysis, with required data-files for the human embryo datasets are available from https://tinyurl.com/mrxnua8k, and with required data-files for the breast cancer at-risk datasets are available from https://tinyurl.com/u8ajrnct

Author contributions

TB conceived of and designed the study. TB wrote the manuscript, with advice from QH. Statistical analyses and model development were carried out by TB, with help from ML.

Funding

Some of the work of TB during this project was funded by the MRC grant MR/P014070/1. The work of QH was supported by the Francis Crick Institute, which receives its core funding from Cancer Research UK (CC2074), the UK Medical Research Council (CC2074), and the Wellcome Trust (CC2074). The funding bodies had no role in the design of the study, collection, analysis, and interpretation of data, or writing of the manuscript.

Acknowledgements

The results published here are in whole or part based upon data generated by The Canadian Epigenetics, Epigenomics, Environment and Health Research Consortium (CEEHRC) initiative funded by the Canadian Institutes of Health Research (CIHR), Genome BC, and Genome Quebec. Information about CEEHRC and the participating investigators and institutions can be found at http://www.cihr-irsc.gc.ca/e/43734.html

The authors acknowledge the use of the UCL High Performance Computing Facilities, and associated support services, in the completion of this work.

Competing interests

The authors declare that there are no competing interests.

License

This manuscript is available under a Creative Commons CC BY-NC License, as described at
https://creativecommons.org/licenses/by-nc/4.0/