Inference of tissue relative proportions of the breast epithelial cell types luminal progenitor, basal, and luminal mature

Thomas E. Bartlett[1*], Swati Chandna[2], and Sandipan Roy[3]

[1]Department of Statistical Science, University College London, U.K.
[2]Department of Economics, Mathematics and Statistics, Birkbeck University of London, U.K.
[3]Department of Mathematical Sciences, University of Bath, U.K.
[*]thomas.bartlett.10@ucl.ac.uk

Abstract

Hormone receptor negative breast cancers are highly aggressive, and are thought to originate from a subtype of epithelial cells called the luminal progenitor. In this paper, we show how to quantify the number of luminal progenitor cells as well as other epithelial subtypes in breast tissue samples using DNA and RNA based measurements. We find elevated levels of these hormone receptor negative luminal progenitor cells in breast tumour biopsies of hormone receptor negative cancers, as well as in healthy breast tissue samples from *BRCA1* (*FANCS*) mutation carriers. We also find that breast tumours from carriers of heterozygous mutations in non-BRCA Fanconi Anaemia pathway genes are much more likely to be hormone receptor negative. These findings have implications for understanding hormone receptor negative breast cancers, and for breast cancer screening in carriers of heterozygous mutations of Fanconi Anaemia pathway genes.

Keywords

Cell type deconvolution; breast cancer; Fanconi Anaemia.

Introduction

Despite advances in diagnosis and treatment in the last several decades, breast cancer remains responsible for more than 5000 deaths per year in the U.K. [1]. Triple-negative breast cancer (TNBC) is a particularly aggressive subtype of breast cancer, with poor prognosis and few treatment options [2]. The triple-negative breast cancer subtype has this name because it lacks receptors for the hormones estrogen and progesterone, as well as the growth-factor HER2. Hence, quantifying hormone receptor negative cells in breast tissue that is at risk of developing cancer is likely to provide important prognostic information.

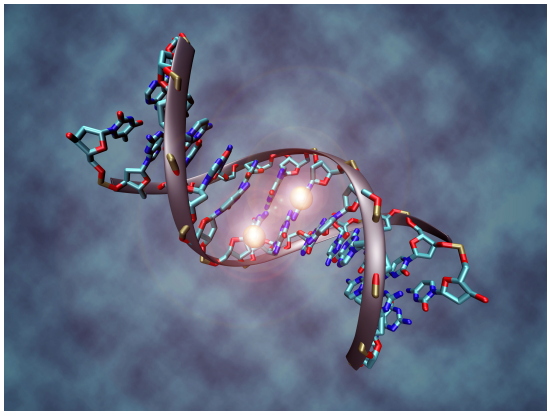

Figure 1: *DNA methylation.*[*]

The cell of origin for TNBC is thought to be a hormone receptor negative epithelial cell called the 'luminal progenitor' cell [3]. Luminal progenitor cells are one of three main cell subtypes of the breast epithelium, along with hormone receptor positive luminal mature cells, and basal epithelial cells: these cells are arranged in luminal and basal layers in the bi-layered epithelium of the breast. Luminal progenitor cells, basal cells, and luminal mature cells are all thought to originate from mammary stem cells, which also reside in the basal layer [4].

DNA methylation (DNAme, Fig.1) is an epigenetic mark that is applied to the DNA at specific cytosine bases, and DNA methylation patterns enable different tissues and cell types to be distinguished from one another [5, 6]. Because DNAme marks are chemical modifications of specific DNA cytosine bases,

[*]Illustration by Christoph Bock, Max Planck Institute for Informatics.

their average levels over a tissue or cell type fluctuate much more slowly than levels of transient mRNA transcripts do. The relative stability, and the tissue and cell type specificity of DNAme marks, along with the observation that variation in mRNA levels may account for less than half of the variation in protein levels [7, 8], means that DNAme data are a promising alternative to gene-expression or RNA-seq data for identifying cell type mixing proportions in bulk-tissue data. DNAme patterns also have a lot of potential as prognostic biomarkers for various cancers, as changes in DNAme levels in a tissue precede formation of cancer at the same site [9]. DNA methylation patterns have been shown to have prognostic power in TNBC [10] as well as other womens' cancers [11, 12]. DNA methylation in bulk-tissue samples is typically analysed in terms of the DNA methylation rate β at each DNA cytosine base. Hence, the methylation rate β represents the fraction of DNA strands in the bulk sample (from several hundred thousand cells) which carry the methylation mark at a specific cytosine locus.

The basic form of the cell type deconvolution model is

$$\mathbf{X} = \mathbf{AW} + \boldsymbol{\epsilon}, \tag{1}$$

where $\mathbf{X} \in \mathbb{R}^{p \times n}$ are the observations on p genomic features in n bulk-tissue (i.e., mixed cell type) samples, $\mathbf{A} \in \mathbb{R}^{p \times k}$ are the genomic profiles over the p features in k cell types, and $\mathbf{W} \in [0,1]^{k \times n}$ are the mixing weights representing the k cell or tissue-type proportions in the n samples. The main object of inference is $\mathbf{W}$, and various methods exist to estimate $\mathbf{W}$ in gene-expression (microarray or RNA-seq) data [13, 14, 15, 16] as well as DNAme data [17, 6]. In practice, $\mathbf{W}$ is estimated by considering a reduced set of p' features ($p' < p$), defined by $\mathbf{A} \in \mathbb{R}^{p' \times k}$.

A key component of any cell type deconvolution algorithm is the identification of the reference genomic profiles $\mathbf{A} \in \mathbb{R}^{p' \times k}$ for the k main constituent cell types in the mixture. This reference matrix $\mathbf{A}$ can be thought of as representing a p'-dimensional space in which the cell types of interest can be distinguished well from one-another. In practice, $\mathbf{A}$ may be determined as part of a unified procedure for fitting the model of Eq.1 [18, 17], or $\mathbf{A}$ may be estimated *a-priori* from external reference data-sets [6, 19, 13, 15, 16]. In this paper we focus on inference of $\mathbf{W}$ using *a-priori* estimates of $\mathbf{A}$. To estimate $\mathbf{A} \in \mathbb{R}^{p' \times k}$, a standard approach is to select p' features which act as 'markers' for the k cell types of interest based on a reference data-set, and then take A_{il} as the mean of the observed values in the reference data-set for cell type $l \in \{1, ..., k\}$ in feature (e.g., gene, or cytosine locus) $i \in \{1, ..., p'\}$. These p' features are typically identified by differential expression analysis for each cell type individually (in comparison with all the other cell types pooled) [20].

In this work, we show how to estimate the concentrations of the breast epithelial cell subtypes luminal progenitor, luminal mature and basal. We do this using both DNA methylation and RNA-seq (i.e., gene-expression) data. We show the robustness of the DNAme-based classifier, and evaluate it against several RNA-based methods. We then show that the DNAme-based breast epithelial subtype classifier reproduces expected proportions of the epithelial subtypes in biopsies of hormone receptor positive and negative tumours. Using the DNAme-based breast epithelial subtype classifier, we explain important differences in inter-individual tumour heterogeneity in terms of variations in luminal progenitor cells that are associated with carriers of heterozygous mutations in non-BRCA Fanconi Anaemia pathway genes.

Results

In this section, we first present the results of a simulation study; we then compare inferences of cell type proportions based on data from DNA and RNA; finally, we present our findings when applying these methods to data from breast tumour biopsies and healthy breast tissue samples.

Simulation study

For the DNAme reference profiles for the breast epithelial subtypes luminal progenitor, luminal mature, and basal, we used previously-published DNAme data from bisulphite-sequencing experiments [21], leading to reference matrix $\mathbf{A} \in [0,1]^{58 \times 3}$ (where $A_{il} \in [0,1]$ rather than $A_{il} \in \mathbb{R}$ because we also have methylation rate $\beta \in [0,1]$, for further details see Methods). To estimate $\mathbf{W}$ (the relative proportions of the breast epithelial subtypes), we used a hierarchical procedure, first estimating the relative proportions of the general cell types epithelial, adipose, stromal, and immune [6], before estimating the proportions of the breast epithelial subtypes. We estimate the tissue relative proportions $W_{:,j}$ of k cell-types in bulk-tissue sample j as:

$$\mathbf{W}_{:,j} = \mathbf{A}^{\dagger}\mathbf{X}_{:,j},$$

where $\dagger$ is a pseudo-inverse of $\mathbf{A}$ calculated via robust partial correlations [6] using the `rlm()` function of the `MASS` package in `R`.

To test how robust the procedure is for estimating concentrations of luminal progenitor, luminal mature and basal epithelial subtypes from DNAme data, we generated simulated data based on Eq.1, together with this pre-defined breast epithelial subtype reference matrix **A**, with $n = 100$ simulated bulk-tissue samples, as follows. After generating $\widetilde{W}_{lj} \sim \mathcal{U}(0,1)$ independently for $l \in \{1,3\}$, and $j \in \{1,100\}$, and using this to calculate the simulated $\widetilde{\mathbf{X}} = \mathbf{A}\widetilde{\mathbf{W}}$, we then replaced some randomly-chosen instances of these simulated data values $\widetilde{X}_{ij}$ (where i and j are independently sampled) with random noise $\widetilde{X}_{ij} \sim \mathcal{U}(0,1)$. We repeated this procedure $b = 1000$ times. Figure 2 shows the results of these robustness tests. We find that with up to 40% of data-values replaced with noise, the results remain very good.

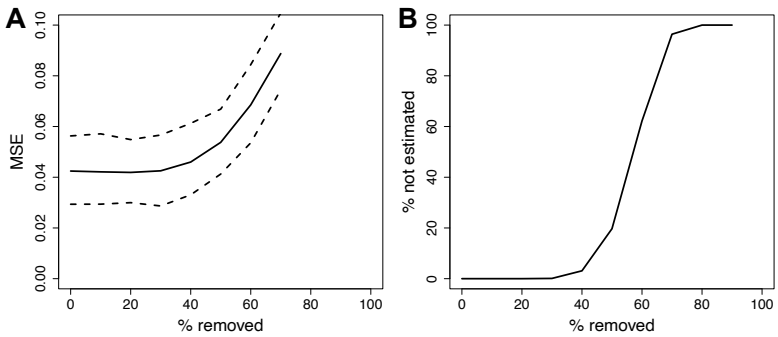

Figure 2: *Model robustness. A data-set with $n = 100$ bulk-tissue samples was simulated, and data-points were removed and replaced at random with noise; this was repeated $b = 1000$ times. (a) Mean square error (MSE) comparing predicted proportions of epithelial subtypes with the ground-truth, as an increasing number of data-points are removed. (b) Percentage of mixing proportions $l \in \{1,k\}$, W_{lj}, $j \in \{1,n\}$, which cannot be estimated, as an increasing number of data-points are removed.*

Comparison of RNA and DNA based inference

For RNA-seq reference profiles for the breast epithelial subtypes luminal progenitor, luminal mature, and basal, we used publicly-available single-cell RNA-seq data for $n = 13909$ breast epithelial cells [4]. We identified which cells corresponded to each breast epithelial subtype, by fitting a Gaussian mixture model (GMM) in the Eigenspace of the graph-Laplacian [22], a procedure we refer to as GMM-LE clustering. Fig.3 shows UMAP (uniform manifold approximation and projection) [23] representations of the data-matrix in two dimensions, illustrating clusters detected by GMM-LE, and average expression levels of independent marker genes for the breast epithelial subtypes luminal progenitor, luminal mature, and basal.

We compared the tissue proportions of the breast epithelial subtypes luminal progenitor, luminal mature and basal, inferred from DNAme with those inferred from RNA-seq data according to a range of existing methods, based on matched data from $n = 175$ low-stage (I-II) breast tumour biopsies downloaded from TCGA (The Cancer Genome Atlas). Table 1 shows correlation coefficients, comparing inferred proportions of the breast epithelial cell subtypes estimated from DNAme data, with those estimated from RNA-seq data according to the URSM [13], Bisque [15], MuSiC [14] and Cibersortx [16] methods. We note that URSM (Unified RNA-Sequencing Model) is the only method achieving correlation close to 0.4 (or greater) when comparing RNA-inferred proportions with DNA-inferred proportions

Method	Luminal progenitor	Luminal mature	Basal
URSM [13]	0.46 (0.33-0.57)	0.71 (0.62-0.77)	0.39 (0.26-0.51)
Bisque [15]	0.45 (0.32-0.56)	0.59 (0.48-0.68)	0.25 (0.1-0.38)
MuSiC [14]	0.57 (0.47-0.67)	0.56 (0.45-0.65)	0.24 (0.1-0.38)
Cibersortx [16]	0.41 (0.27-0.52)	0.24 (0.09-0.37)	0.14 (-0.01-0.28)

Table 1: *Correlation coefficients (with 95% C.I.s), comparing DNAme-estimated proportions of breast epithelial cell subtypes with RNA-seq estimated proportions, according to various methods, in $n = 175$ low-stage breast tumour samples.*

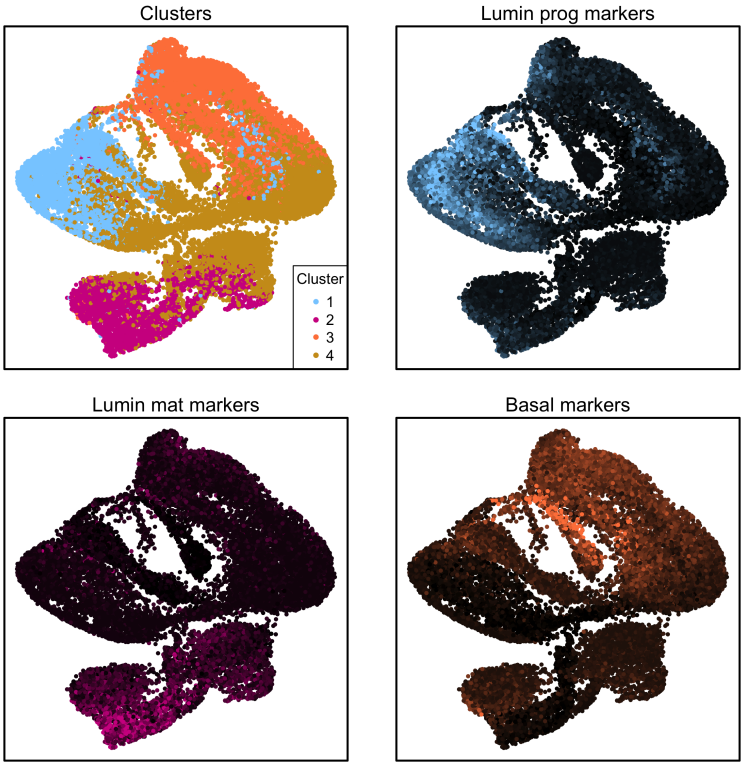

Figure 3: *UMAP projections of single-cell RNA-seq data from $n = 13909$ breast epithelial cells, with (a) inferred GMM-LE clusters, and (b)-(d) colour-intensities of independent and pre-defined sets of marker genes for luminal progenitor, luminal mature, and basal cells (respectively).*

across all cell types (including basal). Noting again that the correlation of mRNA levels with levels of their corresponding proteins is often around 0.4 [7, 8], i.e., when comparing expression levels quantified by different genomic data modalities, we infer that the URSM method for RNA-seq data, as well as the DNAme method, are effective and robust for inferring concentrations of breast epithelial subtypes.

Application to breast cancer biopsies and controls

Luminal progenitor cells are thought to be the cell-of-origin for triple-negative breast cancer (TNBC), which fits well with the model of luminal progenitor cells being similarly hormone receptor negative (HR-) [3]. To test whether the proportions of the breast epithelial subtypes luminal progenitor, luminal mature, and basal (as inferred from DNAme data) are in line with this biological model, we applied the DNAme breast epithelial cell type deconvolution method to DNAme data from TCGA (The Cancer Genome Atlas). Fig.4a shows the concentrations of the breast epithelial subtypes estimated from DNAme data in hormone receptor positive and negative cancers (as classified by expression levels of ESR1, PGR and ERBB2/HER2 in matched gene expression microarray data). As well as elevated levels of luminal progenitors in HR- (hormone receptor negative) breast cancer, Fig.4a also shows elevated level of luminal mature cells (which are HR+) in HR+ breast tumours, also in line with this biological model. To further test whether the proportions of the breast epithelial subtypes inferred from DNAme data are in line with established biological theory, we analysed a publicly-available data-set from GEO (Gene Expression Omnibus). It is well known that the breast epithelial tissue of heterozygous carriers of *BRCA1* mutations contains higher levels of luminal progenitors than would be expected in the general population [24], and this is confirmed in Fig.4b. *BRCA1* (*FANCS*) mutation carriers who unfortunately go on develop breast cancer mostly develop tumours which are HR- [24], presumably due to inefficient DNA damage repair, which is also in line with this model.

The *BRCA1* (*FANCS*) mutation is the best known example of a mutation in the Fanconi Anaemia (FA) / DNA damage repair pathway. Next, we investigated whether similar observations could be made in non-BRCA FA-pathway genes (non-BRCA FANC genes). We used DNAme data from $n = 257$ tu-

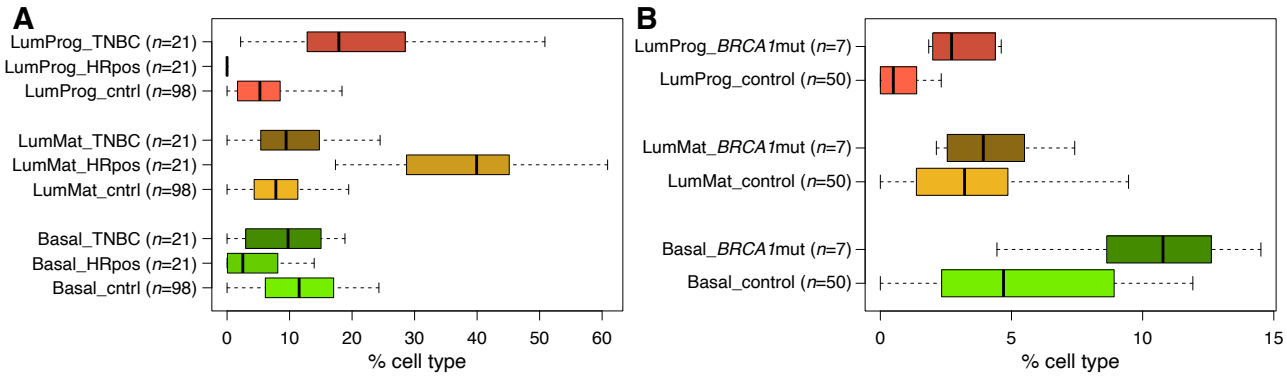

Figure 4: *(a) Inferred proportions of luminal progenitor, basal, and mature luminal cells, in hormone receptor positive / negative cancers. (b) Inferred proportions of luminal progenitor, basal, and mature luminal cells, in BRCA1 (FANCS) mutation carriers.*

mour samples from TCGA with matched clinical data to test how heterozygous carriers of mutations in non-BRCA FANC genes affect the proportions of the breast epithelial subtypes luminal progenitor, luminal mature, and basal cells in tumours that unfortunately arise in this population. To do this, we used multivariate linear models to explain the variation in the concentrations of luminal progenitor, luminal mature, and basal cells in tumour biopsies in terms of non-BRCA FANC gene mutation status, as well as the clinical covariates age and disease stage. We observed that carriers of non-BRCA FANC genes have significantly elevated levels of luminal progenitors and basal cells (Fig.5a-b, significance indicated by C.I. bars), but significantly decreased levels of mature luminal cells, as well as of epithelial cells overall (Fig.5c-d, epithelial proportions calculated as the sum of the proportions of the individual epithelial subtypes). This again fits with the biological model of inefficient DNA damage repair in heterozygous carriers of mutations in FA pathway genes, with this manifesting as elevated levels of cells with replicative potential in the epithelial tissue (i.e., luminal progenitors and basal cells). Presumably this pool of cells with replicative potential is prone to expand more quickly in this population who may have inefficient DNA damage repair mechanisms due to heterozygous mutations in FA pathway genes. This may be compounded by somatic mutations that arise by chance in the other allele of the gene with the FA pathway mutation. This would be expected to correspond to an increase in prevalence of HR- tumours relative to HR+ tumours in FA-gene mutation carriers, as previously reported [25], and this is confirmed in Fig.5e-f. The 10 HR- tumour biopsies from carriers of heterozygous mutations in non-BRCA FA genes represented in Fig.5e correspond to 2 carriers with SNPs in each of FANCD2, FANCF, FANCG, FANCI, and FANCM (one carrier had SNPs in both FANCG and FANCI), as well as one carrier with a SNP in FANCB. Several of these genes have already been associated with breast cancer susceptibility; particularly, FANCD2 [26], FANCG [27, 28], and FANCM [28].

Discussion

We have presented methods to infer proportions of the breast epithelial cell subtypes luminal progenitor, luminal mature, and basal, using DNA and RNA-based methods. We have compared the results using several existing algorithms for estimating the matrix of mixing proportions $\mathbf{W} \in [0,1]^{k \times n}$, for k cell types in n bulk-tissue samples. In particular, we found good agreement (as assessed by Pearson correlation) between the DNAme (DNA methylation) based inference method and the RNA-seq based inference method called URSM [13], that are in line with the previously-reported level of agreement between expression levels of a gene quantified by different genomic data modalities.

Our results when applying these methods to tumour biopsies and healthy breast tissue samples agree well with established biological models of the cell-of-origin for hormone receptor negative breast cancers [3]. Specifically, we found significantly elevated levels of luminal progenitors (which are hormone receptor negative) in biopsies of triple-negative (i.e., hormone receptor negative) breast cancers. Furthermore, we found significantly elevated levels of luminal progenitors in the healthy breast epithelial tissue of heterozygous carriers of *BRCA1* (*FANCS*) mutations, as previously reported [24].

Investigating whether these findings extend to non-BRCA FA-pathway genes (non-BRCA FANC genes), we found that the concentration of luminal progenitors in a tissue sample was significantly

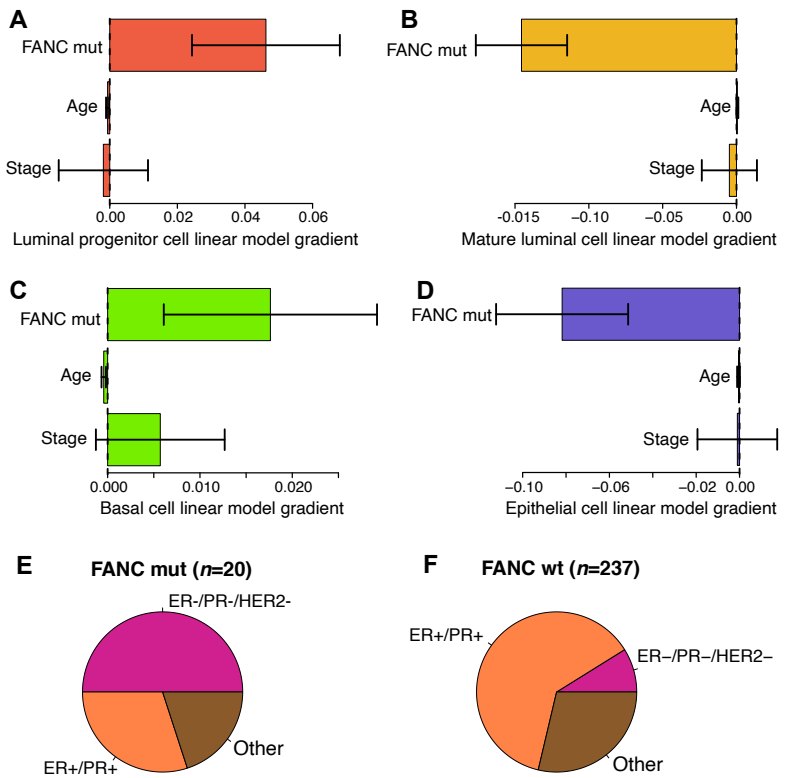

Figure 5: *Linear modelling (bars show 95% C.I.) of (a) luminal progenitor, (b) mature luminal, (c) basal, and (d) overall epithelial cell proportions, predicted by FA-gene (non-BRCA, FANC suffix) mutation carrier status, with disease stage and volunteer age as covariates, in* $n = 257$ *tumour biopsies. (e) Biopsies of FA-gene mutation carriers are mostly hormone receptor (HR) triple-negative, whereas those of (f) FA-gene wild-type are mostly HR-positive.*

predicted by the presence of a heterozygous non-BRCA FANC mutation, in a model that adjusts for the important clinical covariates disease stage and patient age. This is in line with observations based on the same data-set, that incidence of hormone receptor negative tumours make up a significantly much greater proportion of the total number of tumours in heterozygous carriers of non-BRCA FANC mutations (50% compared to 9%). This finding has implications for cancer screening in this population of heterozygous carriers of non-BRCA FANC mutations [29], given the relative much worse prognosis of hormone receptor negative breast tumours compared to breast cancers overall, and may also be relevant to screening for other epithelial cancers in this population.

Methods

Bulk-tissue DNAme microarray data were downloaded from TCGA (`https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`) and GEO (`https://www.ncbi.nlm.nih.gov/geo/`) and pre-processed as follows. Data were background-corrected, probes were removed with $< 95\%$ coverage across samples, and any remaining probes with detection $p > 0.05$ were replaced by k-NN imputation, with $k = 5$. Cell type specific bisulphite-sequenced DNAme data for the breast epithelial subtypes luminal progenitor, luminal mature, and basal were downloaded from EGA (European Genome-phenome Archive), `https://ega-archive.org`) and pre-processed as follows. Sequencer reads were aligned and counted using Bismark [30] with default settings. We subsequently retained only reads mapping to DNA cytosine loci that are also measured in the DNAme microarray data, and which had a total number of mapped reads (methylated+unmethylated) of at least 20 (leading to a granularity of methylation rate of at least 0.05). In this breast epithelial cell type specific data-set, only one reference profile is available for each epithelial subtype. So to select p' features (i.e., DNA cytosine loci) of interest, we selected features according to the following criteria: (1) Low variance of methylation rate β across cell types of non-epithelial lineages (specifically, $\mathbb{V}\beta < 0.001$, where $\mathbb{V}$ is the variance opera-

tor). (2) A difference in mean methylation rate β of at least 0.5 between the epithelial subtype of interest and non-epithelial lineages. (3) The greatest difference in mean methylation rate β between the breast epithelial subtypes. This results in a reference matrix $\mathbf{A} \in [0,1]^{58\times 3}$. To estimate the concentrations of the breast epithelial subtypes in the presence of stromal and adipose cells, we use a hierarchical procedure, first estimating the relative proportions of the general cell types epithelial, adipose, stromal, and immune [6], before estimating the proportions of the breast epithelial subtypes.

Single-cell RNA-seq data for breast epithelial cells were downloaded from GEO (`https://www.ncbi.nlm.nih.gov/geo/`), and libraries with at least 250 reads and transcripts expressed in at least 1000 cells were analysed further. These data were used to define the RNA-seq reference profiles for the breast epithelial subtypes. To estimate the concentrations of these epithelial subtypes in the presence of stromal and adipose cells, we augmented this single-cell breast epithelial data-set with synthetic data based on bulk RNA-seq libraries from purified breast stromal [21] cells and from adipose tissue samples from ENCODE (Encyclopedia of DNA Elements, `https://www.encodeproject.org`). We carried out this augmentation by randomly sampling a single cell library from the breast epithlelial data-set, then replacing the values of the chosen adipose or stromal bulk library with those of the equivalent quantiles of the sampled breast epithelial single cell library, as follows. Define $\widetilde{X}_{\cdot j} = \widehat{F}_b^{-1}\left[\widehat{F}_{ref}(x^{ref})\right]$, where $x^{ref} \in \mathbb{R}^p$ is the vector of (adipose or stromal) reference values across all p genes, $\widehat{F}_{ref}$ is the empirical CDF of x^{ref}, $\widehat{F}_b$, $b \in \{1, ..., B\}$ is the equivalent empirical CDF of a randomly selected epithelial cell, and $\widetilde{X} \in \mathbb{R}^{p\times B}$ is the data-matrix of sampled, normalised values for the adipose or stromal reference to use in the subsequent cell type deconvolution procedure. We then identified the reference matrix $\mathbf{A} \in \mathbb{R}^{p'\times k}$ by identifying the top 100 most significantly upregulated genes for each cell type $l \in \{1, ..., k\}$, when compared to all other cell types. We used the edgeR package in R to test genes for differential expression.

Availability of data and materials

The DNA methylation breast epithelial cell subtype deconvolution tool is made available as an R package from `https://github.com/tombartlett/BreastEpithelialSubtypes`

DNA methylation microarray data for breast cancer and controls were downloaded from TCGA and from GEO under accession number GSE69914. Matched gene-expression microarray data for breast cancer and controls were downloaded from TCGA.

Bisulphite-sequenced DNAme data for the purified breast epithelial cell subtypes luminal progenitor, luminal mature, and basal, were downloaded from the European Genome-Phenome Archive (EGA) under accession number EGAS00001000552

Single-cell RNA-seq data for the breast epithelial cell subtypes luminal progenitor, luminal mature, and basal, were downloaded from the Gene Expression Omnibus (GEO) under accession number GSE113197

Bulk RNA-seq data for breast stromal cells were downloaded from the European Genome-Phenome Archive (EGA) under accession number EGAS00001000552. Bulk RNA-seq data from the adipose tissue samples ENCFF072HRK, ENCFF654JLY, ENCFF732LRY and ENCFF924FNY were downloaded from ENCODE.

References

[1] Breast cancer mortality statistics; 2021. www.cancerresearch.uk/health-professional/cancer-statistics/statistics-by-cancer-type/breast-cancer/mortality.

[2] Hudis CA, Gianni L. Triple-negative breast cancer: an unmet medical need. The oncologist. 2011;16:1–11.

[3] 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.

[4] Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nature communications. 2018;9(1):1–12.

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

[6] 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.

[7] Maier T, Güell M, Serrano L. Correlation of mRNA and protein in complex biological samples. FEBS letters. 2009;583(24):3966–3973.

[8] Schwanhäusser B, Busse D, Li N, Dittmar G, Schuchhardt J, Wolf J, et al. Global quantification of mammalian gene expression control. Nature. 2011;473(7347):337–342.

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

[10] Fackler MJ, Cho S, Cope L, Gabrielson E, Visvanathan K, Wilsbach K, et al. DNA methylation markers predict recurrence-free interval in triple-negative breast cancer. NPJ breast cancer. 2020;6(1):1–6.

[11] 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.

[12] Bartlett TE, Zaikin A, et al. Detection of epigenomic network community oncomarkers. The Annals of Applied Statistics. 2016;10(3):1373–1396.

[13] Zhu L, Lei J, Devlin B, Roeder K. A unified statistical framework for single cell and bulk RNA sequencing data. The annals of applied statistics. 2018;12(1):609.

[14] Wang X, Park J, Susztak K, Zhang NR, Li M. Bulk tissue cell type deconvolution with multi-subject single-cell expression reference. Nature communications. 2019;10(1):1–9.

[15] Jew B, Alvarez M, Rahmani E, Miao Z, Ko A, Garske KM, et al. Accurate estimation of cell composition in bulk expression through robust integration of single-cell information. Nature communications. 2020;11(1):1–11.

[16] Steen CB, Liu CL, Alizadeh AA, Newman AM. Profiling cell type abundance and expression in bulk tissues with CIBERSORTx. In: Stem Cell Transcriptional Networks. Springer; 2020. p. 135–157.

[17] Houseman EA, Molitor J, Marsit CJ. Reference-free cell mixture adjustments in analysis of DNA methylation data. Bioinformatics. 2014;30(10):1431–1439.

[18] Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the national academy of sciences. 2004;101(12):4164–4169.

[19] van Dijk D, Nainys J, Sharma R, Kaithail P, Carr AJ, Moon KR, et al. MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. BioRxiv. 2017;p. 111591.

[20] Newman AM, Liu CL, Green MR, Gentles AJ, Feng W, Xu Y, et al. Robust enumeration of cell subsets from tissue expression profiles. Nature methods. 2015;12(5):453–457.

[21] 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):2060–2074.

[22] Rubin-Delanchy P, Priebe CE, Tang M, Cape J. A statistical interpretation of spectral embedding: the generalised random dot product graph. arXiv preprint arXiv:170905506. 2017;.

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

[24] Lim E, Vaillant F, Wu D, Forrest NC, Pal B, Hart AH, et al. Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nature medicine. 2009;15(8):907–913.

[25] Gordiev M, Brovkina O, Shigapova L, Shagimardanova E, Enikeev R, Nikitin A, et al. Heterozygous mutation in fanconi anemia genes associated with hereditary breast cancer. Annals of Oncology. 2019;30:iii10.

[26] Barroso E, Milne R, Fernandez L, Zamora P, Arias JI, Benítez J, et al. FANCD2 associated with sporadic breast cancer risk. Carcinogenesis. 2006;27(9):1930–1937.

[27] Neidhardt G, Hauke J, Ramser J, Groß E, Gehrig A, Müller CR, et al. Association between loss-of-function mutations within the FANCM gene and early-onset familial breast cancer. JAMA oncology. 2017;3(9):1245–1248.

[28] Schubert S, van Luttikhuizen JL, Auber B, Schmidt G, Hofmann W, Penkert J, et al. The identification of pathogenic variants in BRCA1/2 negative, high risk, hereditary breast and/or ovarian cancer patients: High frequency of FANCM pathogenic variants. International journal of cancer. 2019;144(11):2683–2694.

[29] D'Andrea AD. Susceptibility pathways in Fanconi's anemia and breast cancer. New England Journal of Medicine. 2010;362(20):1909–1919.

[30] Krueger F, Andrews SR. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. bioinformatics. 2011;27(11):1571–1572.

Author contributions

TB conceived and designed the study and wrote the manuscript. All analyses were carried out by TB under advice from SC and SR.

Funding

The work of TB during this project was funded by the MRC grant MR/P014070/1. The funding body had no role in the design of the study, collection, analysis, and interpretation of data, or writing of the manuscript.

Acknowledgements

TB dedicates his work on this manuscript to the memories of Joanne Walker and Joel Walker. The authors are grateful to Peter Kennedy for reading the manuscript and for his feedback which has improved the work. 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`

Competing interests

The authors declare that there are no competing interests.