ARTICLE OPEN

Check for updates

Reclassifying tumour cell cycle activity in terms of its tissue of origin

Arian Lundberg[1,2,3,4], Joan Jong Jing Yi[5], Linda S. Lindström[2] and Nicholas P. Tobin[2]✉

Genomic alterations resulting in loss of control over the cell cycle is a fundamental hallmark of human malignancies. Whilst pan-cancer studies have broadly assessed tumour genomics and their impact on oncogenic pathways, analyses taking the baseline signalling levels in normal tissue into account are lacking. To this end, we aimed to reclassify the cell cycle activity of tumours in terms of their tissue of origin and determine if any common DNA mutations, chromosome arm-level changes or signalling pathways contribute to an increase in baseline corrected cell cycle activity. Combining normal tissue and pan-cancer data from over 13,000 samples we demonstrate that tumours of gynaecological origin show the highest levels of corrected cell cycle activity, partially owing to hormonal signalling and gene expression changes. We also show that normal and tumour tissues can be separated into groups (quadrants) of low/high cell cycle activity and propose the hypothesis of an upper limit on these activity levels in tumours.

npj Precision Oncology (2022)6:59 ; https://doi.org/10.1038/s41698-022-00302-7

INTRODUCTION

The cell cycle is an all-encompassing term that describes the way by which a cell duplicates its genetic contents and subsequently divides into two identical daughter cells. The cycle consists of two main events – interphase and the mitotic or M phase[1]. Interphase can be subdivided into the S phase, where DNA replication occurs, and either side of this come the Gap 1 (G1) and 2 (G2) phases. In addition to the G1, S, G2, and M phases the cell can also enter a quiescent or non-proliferative state after prolonged serum withdrawal termed G0[2]. Reintroduction of serum allows the cell to again enter the cell cycle at G1. Transitions between the cell cycle phases are governed by the cyclin-dependent kinases (CDKs) and their binding partners from the cyclin family of proteins, cyclin D, E, A and B.

Loss of control over the cell cycle resulting in uncontrolled proliferation is a hallmark of cancer[3] and aberrant expression of cell cycle-related genes is frequently observed at a pan-cancer level[4]. The most common cell cycle gene alterations include deletion of the p16 tumour suppressor gene (*CDKN2A/B*, deleted in approximately 20% of all cancer patients), deletions and mutations in the retinoblastoma protein (*RB1*, 7% of patients) and amplifications of the cyclin D1 (*CCND1*, 6% of all cancer patients, up to 30% in breast cancer[5]), E1 (*CCNE1*, 3.6%) and *CDK4* (3.2%) genes[6]. In line with this, pan-cancer studies have also indicated that genomic aberrations are more common in signalling pathways that promote S phase entry or prevent cell cycle exit (e.g. DNA damage response pathway)[4] rather than those associated with mitotic entry and exit (for review see here ref. [1]).

We recently conducted a comprehensive pan-cancer analysis of the genomic aberrations present in tumour samples based on the magnitude of cell cycle pathway activity[7]. We found that cell cycle activity varies broadly across and within the cancer types of The Cancer Genome Atlas (TCGA) pan-cancer cohort and that *TP53*, *PIK3CA* and chromosomal alterations occur with increasing frequency in tumours with increasing cell cycle activity. Here, we build on this work, hypothesising that the starting or baseline level of cell cycle activity/cell proliferation in normal tissue influences the level of tumour cell cycle activity and by extension, that we can reclassify tumour cell cycle activity by placing it in terms of its normal tissue of origin. To test this hypothesis, we analyse over 13,000 samples from the UCSC Toil RNA-seq Recompute Compendium of batch corrected RNA-seq data from the Genotype-Tissue Expression (GTEx, normal tissue) and pan-cancer TCGA projects[8]. Specifically, we apply a gene expression signature representative of the cell cycle (the Cell Cycle Score, CCS[7,9,10]) and strongly correlated to cell proliferation to the samples from the GTEx and TCGA studies and recalculate the cell cycle/proliferative activity of 23 different cancer types taking normal tissue baseline cell cycle levels into account. We also compare the two groups of cancers with the lowest and highest baseline corrected cell cycle activity in order to understand if any specific signalling pathways or genomic aberrations contribute to this increase in cell cycle activity.

RESULTS

GTEx and TCGA samples cluster together on the basis of tissue of origin or cell cycle activity levels

In order to compare cell cycle activity between normal and cancerous tissues, we analysed the GTEx and TCGA pan-cancer datasets which were reprocessed, normalised and batch-corrected together as a part of the Toil recompute project[8]. From the original 19,131 samples we filtered out those that were not a part of these two datasets ($N = 734$), that lacked mRNA-expression data ($N = 92$), that did not have representative tissue from the same site in both studies ($N = 802$) or where the normal tissue or cancer site contained fewer than ten samples ($N = 4043$, Fig. 1). Next, we assessed possible batch effects and the presence of

[1]Department of Radiation Oncology, University of California at San Francisco, San Francisco, CA, USA. [2]Department of Oncology and Pathology, Karolinska Institutet and University Hospital, Stockholm, Sweden. [3]Department of Radiation Oncology, Stanford School of Medicine, Stanford, CA, USA. [4]Helen Diller Family Comperhensive Cancer Center, University of California at San Francisco, San Francisco, CA, USA. [5]School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore.
✉email: nick.tobin@ki.se

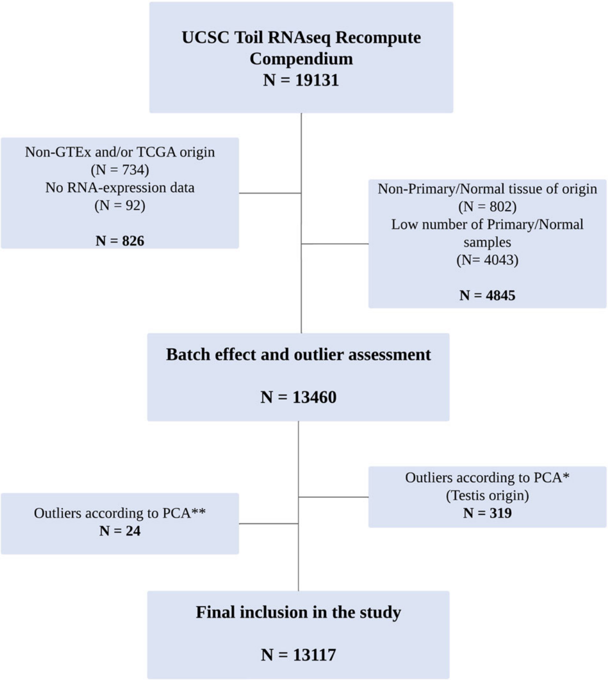

UCSC Toil RNAseq Recompute Compendium
N = 19131
Non-GTEx and/or TCGA origin
(N = 734)
No RNA-expression data
(N = 92)
N = 826
Non-Primary/Normal tissue of origin
(N = 802)
Low number of Primary/Normal samples
(N= 4043)
N = 4845
Batch effect and outlier assessment
N = 13460
Outliers according to PCA*
(Testis origin)
N = 319
Outliers according to PCA**
N = 24
Final inclusion in the study
N = 13117

Fig. 1 CONSORT diagram of sample filtering and selection. *Excluded samples as shown in Supplementary Fig. 1a, b; **excluded samples as shown in Supplementary Fig. 1c, d.

outliers with PCA plots using RNA-seq data from the remaining 13,460 samples. Plotting on the basis of the most variable genes across all samples, PCA showed a long tail of samples stretching into the lower right quadrant (Supplementary Fig. 1a coloured in red, and Supplementary Fig. 1b). Closer inspection and annotation of these samples showed all were GTEx normal testicular tissue (without the presence of TCGA testicular tumours), indicating a potential study of origin batch effect and as such, all normal and cancer testicular samples were removed from further analyses ($N = 319$). Similarly, after applying our Cell Cycle Score (CCS) signature to the remaining samples, we noted a small number of outliers ($N = 24$) from mixed origin using PCA (Supplementary Fig. 1c, d, circled in red), which were also removed from further analysis leaving 13,117 samples in total. A breakdown of the origin and number of all removed samples is shown in Supplementary Table 1.

Replotting the PCA of most varying genes in 13,117 samples ($N = 4979$ normals and 8138 cancers) did, in general, not show any clear grouping on the basis of study origin (Fig. 2a; grey = GTEx, light orange = TCGA normal tissue and blue = TCGA tumours. GTEx and TCGA normals only shown in Supplementary Fig. 2a) but rather on the basis of tissue type (Fig. 2b and Supplementary Fig. 2b). If a study of origin batch effect had been present here we would've expected to see the GTEx and TCGA samples clustering separately to each other without overlap regardless of tissue type. These results indicate an absence of batch effects related to study origin and support further direct comparison of tissue types. A similar finding was observed using UMAP (Supplementary Fig. 1e, f), a method that can find non-linear relationships in data that could be otherwise missed using PCA alone.

We subsequently examined how samples cluster on the basis of the CCS gene signature, and given its strong correlation to cell proliferation, we anticipated that slower growing normal samples would cluster separately from faster growing tumour samples. This was generally the case with normal samples from both GTEx and TCGA clustering on one side of the plot (Fig. 2c, left hand side, grey and light orange dots) and tumours clustering separately beside them (Fig. 2c, blue dots). However, a group of normal samples is also clearly present when the blue TCGA tumour group overlay is removed (Supplementary Fig. 2c). This group contains normal tissues (e.g. skin and oesophageal tissues Supplementary Fig. 2d) that appear to grow as quickly as slowly growing cancer samples (e.g. kidney, Fig. 2d, red dots). Again, in order assess non-linear effects in the data that might not be readily apparent using PCA, we also performed a UMAP analysis using CCS genes and examined clusters on the basis of study origin, tissue type and the CCS signature (Fig. 2e–g). Slower growing normal samples of mainly GTEx origin clustered around the outer edges of the plot (Fig. 2e, grey dots and Fig. 2g, lower CCS = dark blue) and a core group of samples high in CCS was also readily apparent (Fig. 2g, higher CCS = yellow). This CCS high group was mainly comprised of oesophageal, head and neck and cervical cancers (Fig. 2f, g). Some tissue specific clustering of normal and tumour samples was also found, in particular for brain and kidney samples (Fig. 2f). In line with our PCA findings, kidney tumour and normal samples cluster closer to the outer edges of the plot than the high CCS core, implying a generally slower growing tissue type (Fig. 2e, f). Together, these results indicate that GTEx and TCGA samples cluster together on the basis of tissue of origin (as expected, given the aim of the Toil recompute project) but also group on the basis of their cell cycle activity levels which again was expected given the sustained proliferation hallmark of cancer tissues.

Tumours from gynaecological tissues show the highest baseline corrected cell cycle activity levels

To visualise the differences more clearly in cell cycle activity between tumours and their normal tissue of origin, we plotted the CCS as a continuous variable using boxplots and separated samples into tissue of origin. While normal samples showed varying CCS levels when compared to each other (Fig. 3a and Supplementary Fig. 3a), tumour samples had a consistently higher CCS relative to their tissue of origin (Fig. 3a, $P < 0.001$, Students' T-test, adjusted for multiple testing, first boxplot within each tissue type is normal tissue and the rest are tumours). The importance of relating tumour cell cycle activity to normal tissue is also evident from these boxplots: Bladder cancer ("BLCA", red arrow) has higher cell cycle activity than glioblastoma multiforme ("GBM", blue arrow), but normal levels of bladder cell cycle activity are much higher than that of brain tissue (compare "Bladder" to "Brain", black arrows). This means that at the absolute level BLCA shows higher cell cycle activity levels than GBM, but at a relative level (relative to its baseline level in normal tissue) GBM's cell cycle activity is much higher than BLCA. Next, we extended this concept by reclassifying all tumour samples taking (subtracting) the baseline cell cycle activity of its normal tissue of origin into account to give a new Baseline Corrected-Cell Cycle Score (BC-CCS). As the tumour epithelial cell content differs for different cancer types (Supplementary Table 2), we also adjusted the BC-CCS for tumour purity using values derived from the ABSOLUTE algorithm[11]. Notably, we found that in general, cancers of gynaecological origin (Cervical squamous cell carcinoma and endocervical adenocarcinoma, CESC; Ovarian serous cystadenocarcinoma, OV; and Uterine Carcinosarcoma, UCS) display the highest BC-CCS levels and Head and Neck squamous cell carcinoma (HNSC), Kidney Chromophobe (KICH) and Kidney renal papillary cell carcinoma (KIRP) the lowest (Fig. 3b). This finding is even more interesting when placed in the context of our previous research in the TCGA pan-cancer atlas tumours only without baseline correction, where we found HNSC to be a tumour type with one of the highest CCS activity levels[7]. We show here that head and neck tissue also has the highest level of cell cycle activity in normal tissue (Supplementary Fig. 3a). Taken together, this may imply that cell cycle activity in head and neck cancers simply cannot be pushed much higher as its baseline levels are already so

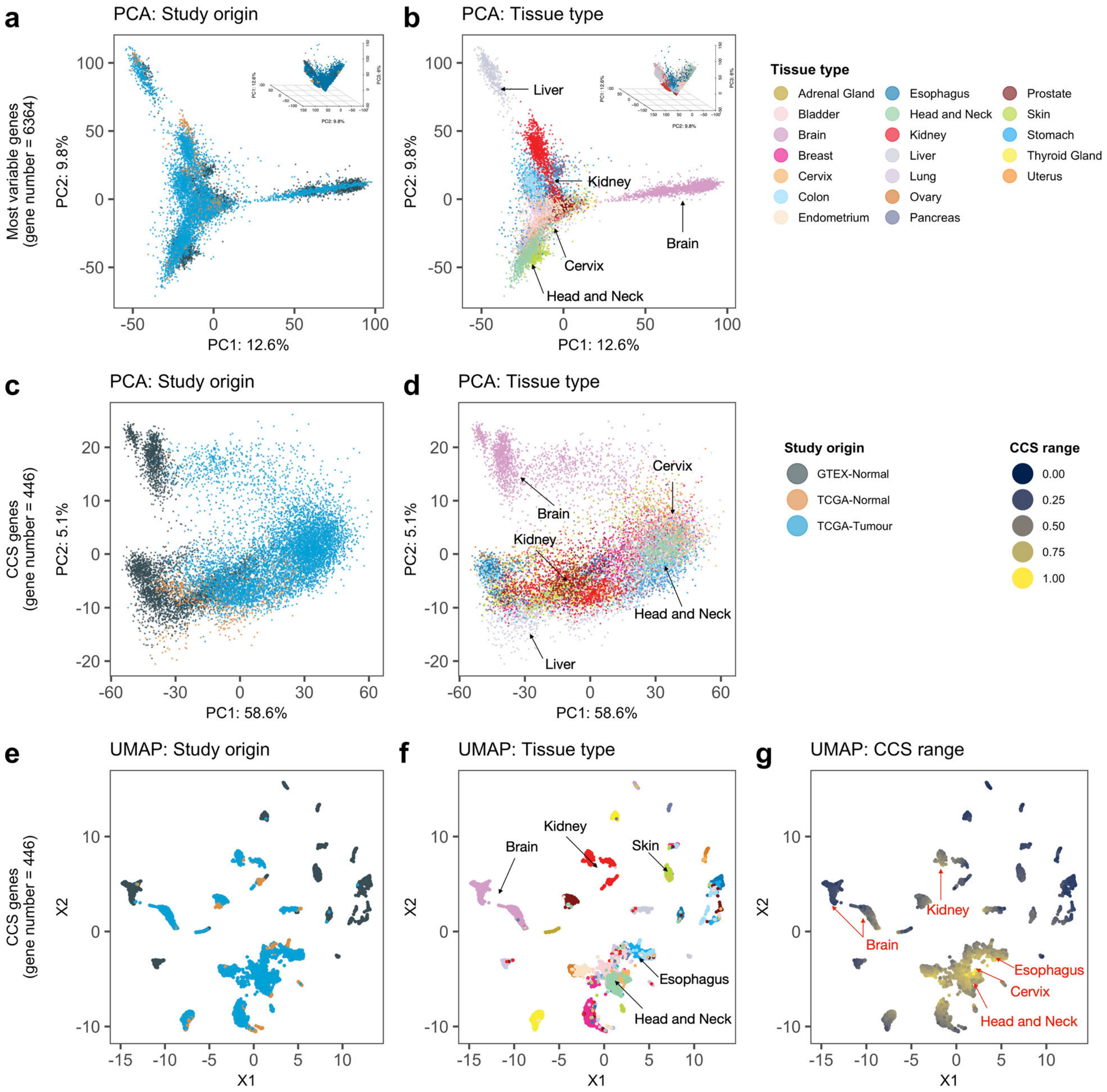


Fig. 2 PCA and UMAP dimensionality reduction of GTEx and Pan-Cancer Atlas datasets. Toil pipeline batch-corrected mRNA expression data from the GTEx and TCGA pan-cancer datasets ($N = 13117$ in total) were used to perform. **a** Principle component analyses (PCA) using the most variable genes in the dataset with an overlay of colours based on study origin or **b** tissue type; **c** PCA using the genes of the CCS signature with an overlay of colours based on study origin or **d** tissue type; **e** Uniform Manifold Approximation and Projection (UMAP) using the genes of the CCS signature with an overlay of colours based on study origin or **f** tissue type or **g** CCS range.

high in normal tissue that it has reached its upper limit or "ceiling". Conversely, gynaecological cancers start at a much lower baseline level (Supplementary Fig. 3a, see Uterus, Ovary, Cervix), giving them a higher ceiling to proliferate into once an oncogenic event has occurred. The relationship between cell cycle activity in normal and tumour tissues based on median CCS expression is visualised in a scatterplot divided into quadrants of high/low CCS in Fig. 3c. Here, bladder, endometrial and head and neck tissues all show high normal/high tumour cell cycle activity, whereas cervix, oesophagus, uterus and ovary tissue show low normal/ high tumour cell cycle activity. The BC-CCS is also visualised anatomically in Supplementary Fig. 4 and similar pan-cancer boxplots where breast cancer is separated into its molecular subtypes are also shown in Supplementary Fig. 3b, c.

Statistically significant differences in DNA copy number when comparing the highest and lowest BC-CCS tumour groups

Next, we wanted to understand if specific signalling pathways or genomic aberrations are more common in tumour types showing a higher BC-CCS compared to those showing lower. For this, we selected the tumour types at the extremes of Fig. 3b, forming two comparison groups (Fig. 3b, red and blue arrows). Group 1 consisted of tumour types showing the lowest BC-CCS (HNSC, KICH and KIRP) with the addition of Uterine Corpus Endometrial Carcinoma (UCEC). UCEC is an exception to our findings in Fig. 3b in that it is a gynaecological tumour type that shows a low BC-CCS. Given that the three highest BC-CCS tumour groups that form Group 2 (CESC, OV, UCS) are all of gynaecological origin, we wanted to make sure that any potential differences found when comparing Group 1 to Group 2 were not only due to a comparison

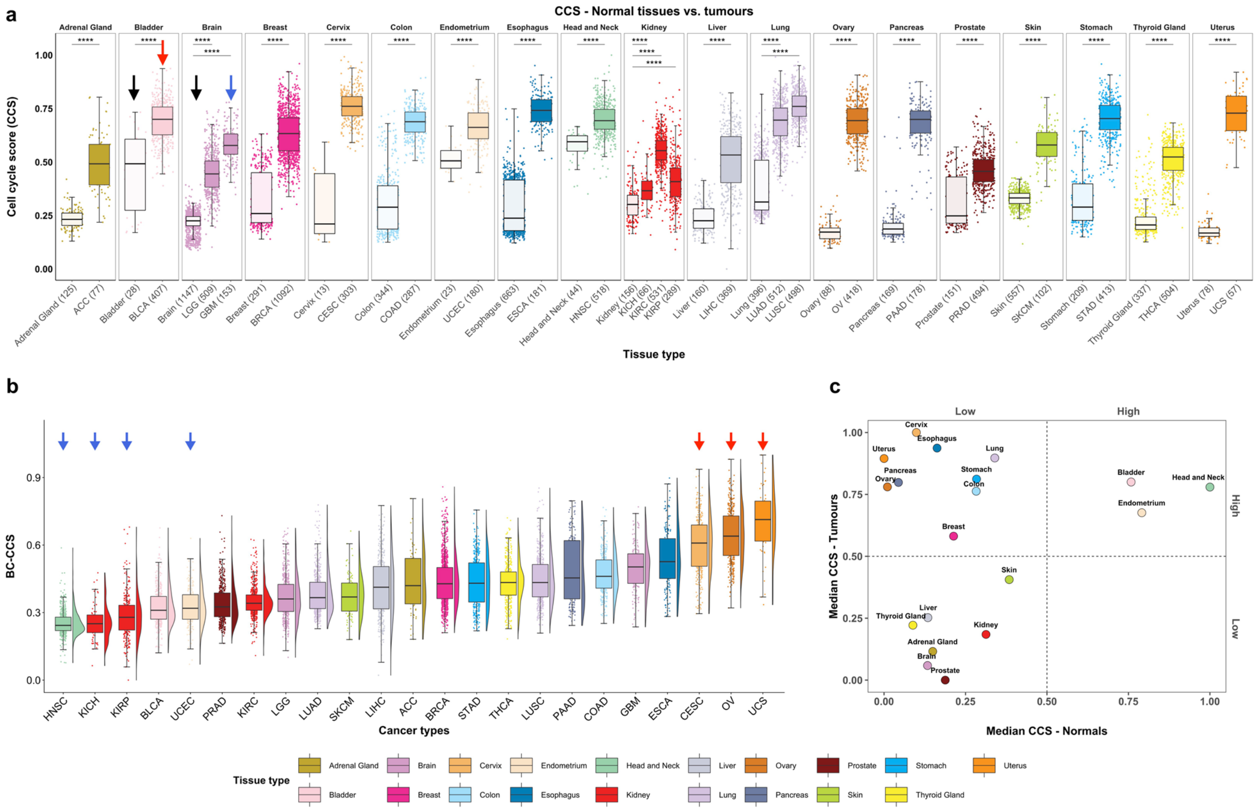


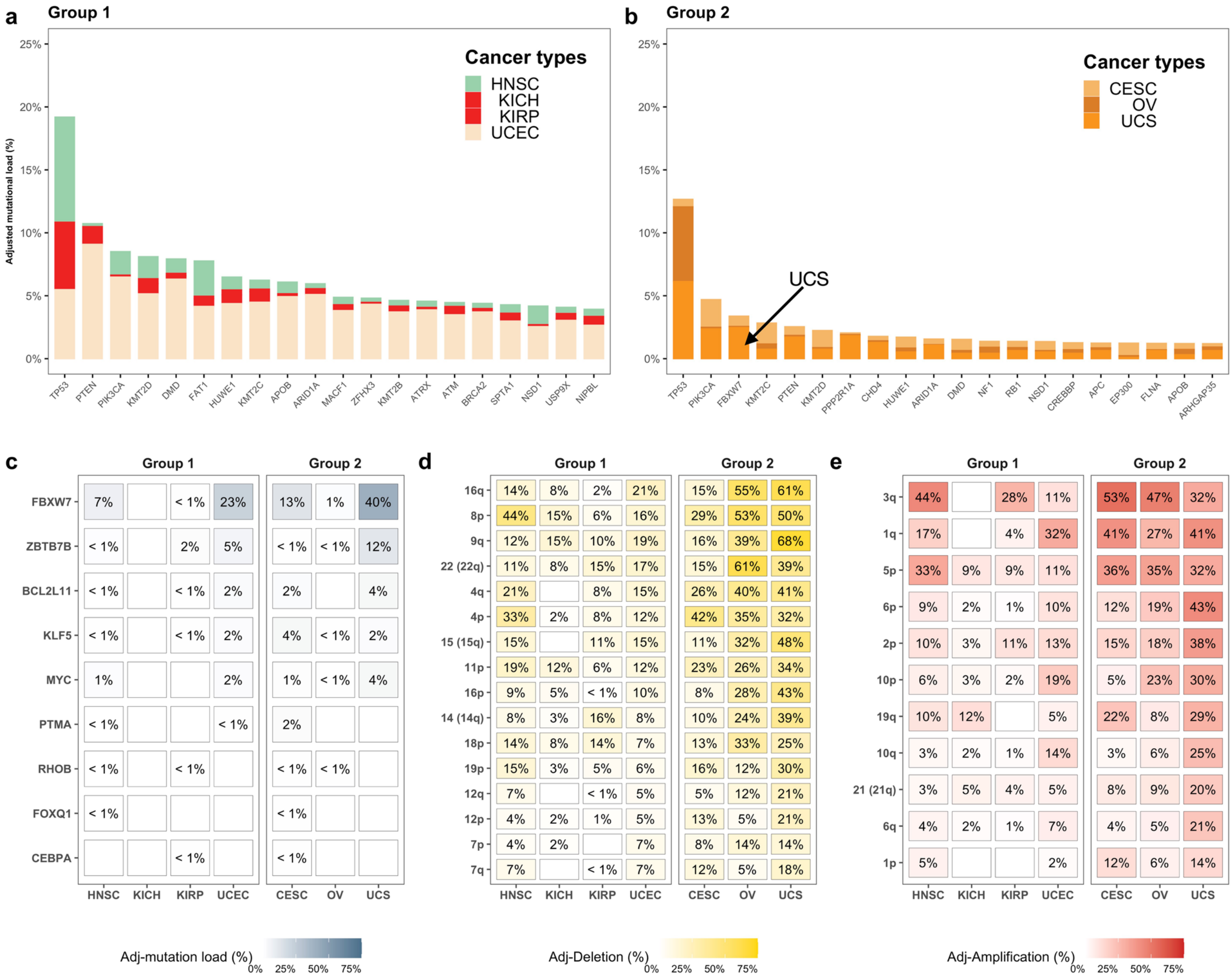
Fig. 3 Boxplots and scatterplot comparing cell cycle activity across normal tissue and Pan-Cancer Atlas tumour samples. Toil pipeline batch corrected mRNA expression data from the GTEx and TCGA pan-cancer datasets were used to examine cell cycle activity. **a** The CCS gene expression signature was used to create boxplots within in each normal and cancer tissue type, black arrows denote normal bladder and brain tissue, red= BLCA, blue = GBM. **b** The baseline corrected-CCS (BC-CCS) in pan-cancer tumours, where the median CCS value for the normal tissue of origin has been subtracted from each cancer type. Density plot is placed beside each boxplot. Blue and red arrows indicate cancers from Groups 1 and 2, respectively. **c** Scatterplot showing the relationship of cell cycle activity in normal tissues with tumour samples on the basis of median CCS expression. Median CCS expression standardised by scaling between 0 and 1. *p* values in boxplots are based on Student's T-test and corrected for multiple testing; **** = $p < 0.0001$. Within each box, horizontal lines denote median values; boxes extend from the 25th to the 75th percentile of each group's distribution of values; vertical extending lines denote adjacent values (the most extreme values within 1.5 interquartile range of the 25th to the 75th percentile of each group).

of gynaecological vs. non-gynaecological tissues. For this reason, we included UCEC in Group 1 as an internal control.

First, we focused on a comparison of DNA-level mutations in 299 known cancer driver genes[12] between these two groups. The top 20 mutations after adjusting for number in each group are shown in Fig. 4a, b and, as anticipated given their high pan-cancer mutation frequency, the *TP53*, *PIK3CA*, and *PTEN* genes are amongst the most mutated in both groups. Using a Fisher's-exact test to assess whether any mutations were statistically more common in Group 2, we found that majority of differences at an individual gene level occurred in a low percentage of tumours (< 5%) within each cancer type (Fig. 4c). One exception to this was the *FBXW7* gene which was mutated in 54% of tumours in Group 2 vs. 31% in Group 1, however, this result is mostly driven by the UCS in Group 2 where 40% of tumour carried this mutation. Second, and continuing our DNA-level analysis, we assessed chromosomal arm-level copy number differences between the same groups. After adjusting for tumour numbers within each cancer type, we found increased deletions in 16 different chromosomal arms in Group 2 relative to Group 1 (Fig. 4d, Fisher's-exact test, FDR < 5% for all comparisons). The largest differences in deletions were found in arms 16q, 8p, 9q, 22q, and 4q (Fig. 4d) and similarly, of the 11 different chromosomal arms found to have significantly increased amplifications between the two groups, the largest differences were found in 3q, 1q, 5p, 6p and 2p (Fig. 4e, Fisher's-exact test and FDR < 5% for all comparisons). As expected, semi-supervised clustering on the basis of these amplifications and deletions shows a reasonable separation of the tumours of Groups 1 and 2 where two BC-CCS groups of low and high activity are discernible (Supplementary Fig. 5, left hand side, BC-CCS variable indicated with red arrow). However, using the same aberrations to cluster all other pan-cancer samples fails to show any clear separation into subgroups with lower or higher BC-CCS (Supplementary Fig. 5, right hand side, BC-CCS indicated with red arrow). This implies that while these arm-level genomic alterations can separate group 1 and 2 tumours on the basis of BC-CCS, they cannot do so at a pan-cancer level and as such are unlikely to more broadly drive the differences in baseline-corrected tumour cell cycle activity at a pan-cancer level.

Statistically significant differences in hormone signalling and gene expression when comparing the highest and lowest BC-CCS tumour groups

As the sex hormones are known mitogens with an established role in pan-gyn studies[13] and in driving the cell cycle, we next determined whether the mRNA levels of the oestrogen receptor alpha (ER-α), progesterone (PR) or androgen receptor (AR) genes were different between Group 1 and Group 2. We found that genes and gene modules (groups of genes) representative of all



c

	Group 1				Group 2		
	HNSC	KICH	KIRP	UCEC	CESC	OV	UCS
FBXW7	7%		< 1%	23%	13%	1%	40%
ZBTB7B	< 1%		2%	5%	< 1%	< 1%	12%
BCL2L11	< 1%		< 1%	2%	2%		4%
KLF5	< 1%		< 1%	2%	4%	< 1%	2%
MYC	1%			2%	1%	< 1%	4%
PTMA	< 1%			< 1%	2%		
RHOB	< 1%		< 1%		< 1%	< 1%	
FOXQ1	< 1%				< 1%		
CEBPA			< 1%		< 1%		

d

	Group 1				Group 2		
	HNSC	KICH	KIRP	UCEC	CESC	OV	UCS
16q	14%	8%	2%	21%	15%	55%	61%
8p	44%	15%	6%	16%	29%	53%	50%
9q	12%	15%	10%	19%	16%	39%	68%
22 (22q)	11%	8%	15%	17%	15%	61%	39%
4q	21%		8%	15%	26%	40%	41%
4p	33%	2%	8%	12%	42%	35%	32%
15 (15q)	15%		11%	15%	11%	32%	48%
11p	19%	12%	6%	12%	23%	26%	34%
16p	9%	5%	< 1%	10%	8%	28%	43%
14 (14q)	8%	3%	16%	8%	10%	24%	39%
18p	14%	8%	14%	7%	13%	33%	25%
19p	15%	3%	5%	6%	16%	12%	30%
12q	7%		< 1%	5%	5%	12%	21%
12p	4%	2%	1%	5%	13%	5%	21%
7p	4%	2%		7%	8%	14%	14%
7q	7%		< 1%	7%	12%	5%	18%

e

	Group 1				Group 2		
	HNSC	KICH	KIRP	UCEC	CESC	OV	UCS
3q	44%		28%	11%	53%	47%	32%
1q	17%		4%	32%	41%	27%	41%
5p	33%	9%	9%	11%	36%	35%	32%
6p	9%	2%	1%	10%	12%	19%	43%
2p	10%	3%	11%	13%	15%	18%	38%
10p	6%	3%	2%	19%	5%	23%	30%
19q	10%	12%		5%	22%	8%	29%
10q	3%	2%	1%	14%	3%	6%	25%
21 (21q)	3%	5%	4%	5%	8%	9%	20%
6q	4%	2%	1%	7%	4%	5%	21%
1p	5%			2%	12%	6%	14%

Fig. 4 Most highly mutated genes and chromosomal arm-level aberrations in Groups 1 and 2. The cancer types with the highest and lowest BC-CCS were placed into two groups: Group 1 = HNSC, KICH, KIRP and UCEC; Group 2 = CESC, OV and UCS. **a** The top 20 most mutated cancer driver genes for Group 1 and **b** Group 2, mutation percentage has been adjusted for sample number within each cancer type for all genes individually. **c** Driver genes with a higher mutation level in Group 2 relative to Group 1 by Fisher's-exact Test. **d** Chromosomal arm-level deletions that occur with higher frequency in Group 2 relative to Group 1 and **e** chromosomal arm-level amplifications occurring with higher frequency in Group 2 relative to Group 1. All comparisons were adjusted for multiple testing.

three hormones show higher levels in Group 2 (Fig. 5a, Student's T-test). When considering all pan-cancer samples, however, the correlation between the three gene modules and the BC-CCS was very weak (see inset of Supplementary Fig. 6, $R = 0.047$, 0.095, and 0.083 for the oestrogen, progesterone and androgen gene modules respective, Spearman's correlation). In addition, boxplots of the same gene modules in individual cancer types show that not all cancer types with a high BC-CCS have high expression of sex hormones (Supplementary Fig. 6a–c). These findings imply that while sex hormone expression may partially explain the increase in BC-CCS in some tumour types, it does not account for a high BC-CCS across all cancers. This is to be expected as sex hormone exposure has not been shown to be a risk factor for all cancer types.

To further examine the genes and pathways altered between Groups 1 and 2, we performed differential gene expression analysis, adjusting for both cancer type and the *ESR1* (ER-α) gene. The latter was included in an attempt to mitigate finding genes that were only representative of the sex hormone differences identified in Fig. 5a. 522 genes were upregulated and 876 downregulated in Group 2 relative to Group 1 (Fig. 5b). Pathway analysis of these genes showed upregulation of the p53, TNFα and, despite adjusting for the *ESR1* gene, oestrogen response pathways (Fig. 5c, Dark blue = significant at 10% FDR). We also examined individual correlations of the differentially expressed genes to the BC-CCS in all tumours of Groups 1 and Group 2 together to determine if any genes were driving the difference in BC-CCS. *C19orf57*, *WT1*, and *RAD9B* were the top 3 most positively and *KCNJ15*, *DNASE1L3* and *NEFL* the most negatively correlated to BC-CCS (Fig. 5d, Spearman's rank correlation). When correlating the same genes to BC-CCS in the rest of the pan-cancer cohort only *C19orf57* showed a moderate correlation (Fig. 5d, Spearman's *Rho* = 0.32). Together, these results again point to a strong influence of oestrogen signalling on the BC-CCS in gynaecological cancers and a potential role for the *C19orf57* gene in driving BC-CCS at a pan-cancer level.

DISCUSSION

In this study, we utilise RNA-seq data from two large public databases of normal and tumour tissue in order to reclassify cancer cell cycle activity in terms of its tissue of origin in over 13,000 samples. Following reclassification, we assessed the DNA mutation, chromosomal copy number and biological pathway

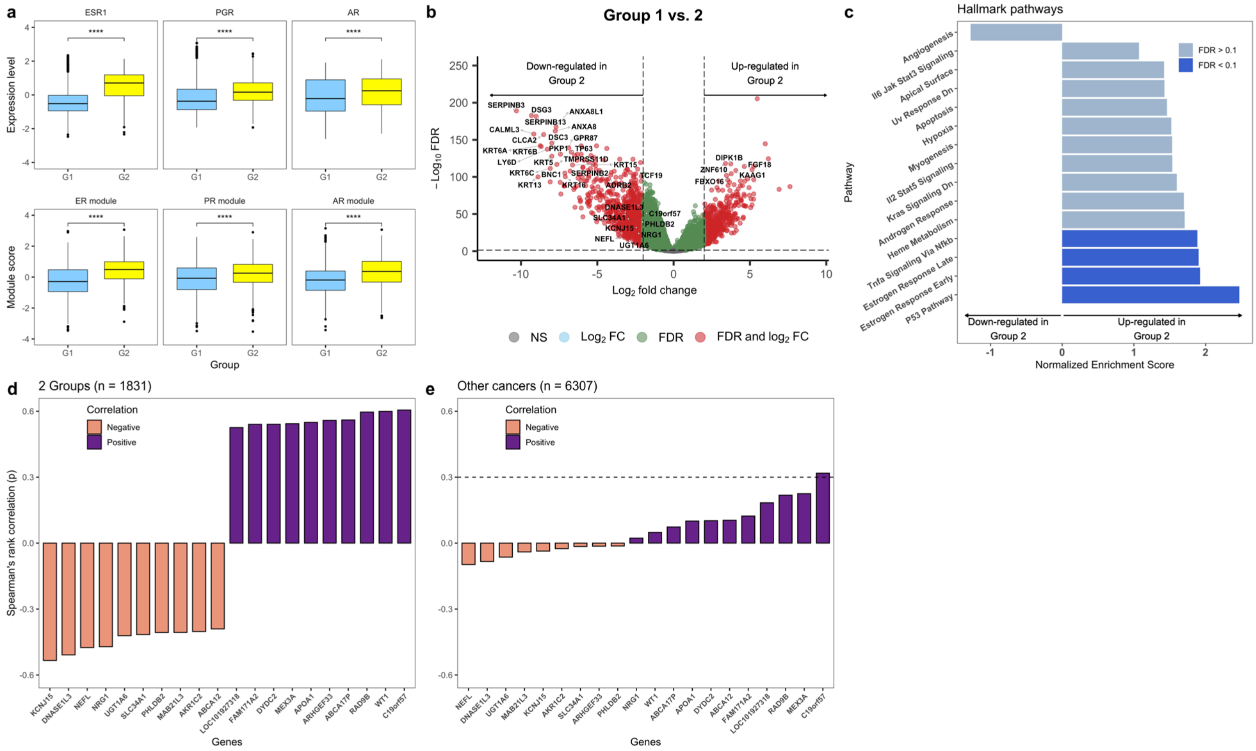


Fig. 5 Comparison of sex hormone gene modules, differentially expressed genes and signalling pathways between Groups 1 and 2. mRNA data from the tumours of Groups 1 and 2 was used to compare: **a** Expression of ER-α, PR and AR genes and representative gene modules. **b** Differential gene expression between the two groups, a volcano plot representing genes that are up- or downregulated in Group 2 relative to Group 1 is shown. **c** Gene set enrichment analyses (GSEA) on the basis of the differentially expressed genes in Group 2. **d** Barplots showing the genes with the highest Spearman's rank correlation to the baseline corrected-Cell Cycle Score (BC-CCS) in the tumours of Groups 1 and 2 only ($N = 1831$) and **e** barplots showing the genes with the highest Spearman's rank correlation to BC-CCS in all other cancer types of the pan-cancer cohort (minus the tumours from Groups 1 and 2, $N = 6307$). *p* values in boxplots are based on Student's T-test and, were corrected for multiple testing; **** $= p < 0.0001$. In the volcano plot genes with FDR < 5% and $\log_2$ fold-change > 2 or <−2 were considered significant. GSEA results with FDR < 10% were considered significant. Spearman's rank correlation *Rho* was used for the correlation barplots; dashed line indicates *Rho* > 0.3. Within each box, horizontal lines denote median values; boxes extend from the 25th to the 75th percentile of each group's distribution of values; vertical extending lines denote adjacent values (the most extreme values within 1.5 interquartile range of the 25th and 75th percentile of each group).

signalling differences between tumours with the largest and smallest baseline corrected cell cycle changes, focusing on tumours of gynaecological origin. Our findings show first that samples grouped broadly together on the basis of their tissue of origin or cell cycle activity level regardless of whether they originated from the GTEx or TCGA datasets. Second, while normal samples showed varying CCS levels when compared to each other, tumour samples had a consistently higher CCS relative to their tissue of origin. Third, in general, tumours of gynaecological origin (CESC, OV, and UCS) show the highest baseline corrected cell cycle change. Fourth, that chromosomal arm-level alterations, hormone receptor signalling and specific genes, including *C19orf57* are associated with this high cell cycle activity in gynaecological tumours and finally, that *C19orf57* also shows a moderate association with BC-CCS at a pan-cancer level.

We and others have previously shown that applying the CCS[7] or cell cycle/proliferation-related classifiers[14] to the TCGA pan-cancer dataset separates tumours into those with high (Testicular Germ Cell tumours – TCGT, Head and Neck squamous cell carcinoma – HNSC and Cervical squamous cell carcinoma and endocervical adenocarcinoma – CESC) and low (Kidney Chromophobe – KICH, Kidney renal papillary cell carcinoma – KIRC, Thyroid carcinoma – THCA) cell cycle activity. Here, we demonstrate that by placing tumour cell cycle activity in terms of its normal tissue of origin, we gain a much clearer understanding of the genomic aberrations and biological signalling pathways that drive the largest changes in cell cycle activity. Indeed, it is striking that the gynaecological tissue types showing the lowest levels of cell cycle activity in normal tissue (OV, UCS - Supplementary Fig. 3a) show the highest baseline corrected levels when examining their oncogenic counterpart. Conversely, head and neck cancers (HNSC) that show the highest levels of cell cycle activity in normal tissue, show the lowest baseline corrected cell cycle levels in tumours. This raises the intriguing question of whether there is a ceiling on cell cycle activity. Is it the case that driver gene mutations in HNSC are unable to push the cell cycle limits further as cells are already cycling at close to their maximum level? By extension, OV and UCS start from such a low level of activity that genomic aberrations (or exposure to sex hormones) have the scope to push cell cycle activity much higher than its starting point. If cancer cells do have an upper limit on their cell cycle or proliferative capabilities, what is the limiting factor? One could speculate that cellular plasticity may hold the answer to these questions. Recent evidence has recognised cellular plasticity or phenotype switching as an essential process in disease (for a review of cell plasticity in cancer cells see here ref. [15]), and one which takes many forms in cancer, including epithelial to mesenchymal (EMT) transition[16], dedifferentiation[17], and transient spatial organisation[18]. While the experimental assessment of these processes lies beyond the scope of the work described herein, Malta et al. provide some evidence of the link between the cell cycle and cellular plasticity at a pan-cancer level. Using machine learning-derived stemness indices that were highly correlated to EMT markers, they showed that higher indices were associated with more proliferative breast cancer subgroups and with head and neck cancers relative to lower indices in kidney renal papillary cell carcinoma[19]. It should be noted, however, that CESC, OV and UCS were of intermediate stemness, implying that stemness is not the only factor relevant to a theoretical upper ceiling of cell cycle activity.

Our analysis aimed to determine if any common genomic aberrations or signalling pathways contribute to an increase in baseline-corrected cell cycle activity. While we did identify clear differences in our Group 1 vs, Group 2 analyses, no aberrations or pathways were strongly associated with BC-CCS at a pan-cancer level. There are likely a number of reasons for this, including signalling pathways that only influence a subset of tumours types (e.g. those responsive to hormonal signalling) or different mutations/ amplifications/ deletions that increase cell cycle activity in a similar way but in different tumour types (e.g. cyclin D1 amplifications in Oesophageal carcinoma and cyclin E1 amplifications in Uterine Carcinosarcoma). We did, however, find a moderate to strong correlation (Spearman's Rho = 0.61) between the BC-CCS and the *C19orf57* gene in Groups 1 and 2, and a moderate one (Spearman's Rho = 0.32) in the rest of the pan-cancer cohort. Also called break repair meiotic recombinase recruitment factor 1 (*BRME1*), this gene has been shown by Zhang et al. to impair the mitotic *BRCA2-RAD51* homologous repair (HR) function in cancer cells and to be upregulated in brain and cervical cancers relative to paired normal tissues[20]. Based on these findings, it could be speculated that through upregulation of *C19orf57* and subsequent sequestration of *BRCA2*, tumours promote genome instability and loss of strict control over the cell cycle. In this case, *C19orf57* gene expression could serve as potential biomarker for impaired HR function and by extension, of patients who may benefit from poly (ADP-ribose) polymerase (PARP) inhibitor treatment. Of relevance, outside of the better-known applications of PARP inhibitors in the gynaecological cancers[21] with the highest BC-CCS (OV, UCEC, CESC), a role of PARP inhibition has also been tested in oesophageal cancer (ESCA)[22] and glioblastoma (GBM)[23]—the tumours types with the fourth and fifth highest BC-CCS (see Fig. 2b). The strength of the correlations we saw in the Group 1 vs. 2 analysis relative to the one observed in rest of the cohort however suggests that *C19orf57* gene expression is unlikely to be the only factor driving large differences in baseline-corrected cell cycle activity. Relatedly, we also saw a number of statistically significant differences in chromosomal arm-level amplifications and deletions in our Group 1 vs. 2 analyses that could be candidates also driving relative cell cycle activity. Their role at a pan-cancer level is less clear though as clustering on their basis showed no pattern of separation into low or high BC-CCS (Supplementary Fig. 5). A second more general implementation of our findings in a precision medicine setting could be as a communication tool to more effectively convey to a patient how quickly their tumour is growing relative to the surrounding tissue. For example, we could have reclassified tumours CCS as fold change relative to normal tissue in this study—this would allow a clinician to explain to a patient that the cells in their tumour are growing e.g. three times more quickly than the surrounding normal tissue and as such it is necessary to employ a more aggressive treatment strategy such as chemotherapy. In practise, however, this would require biopsies or fine needle aspirates from both tumour and normal tissues, likely rendering it unfeasible for implementation.

The CCS gene signature applied here has been derived on the basis of genes known to be expressed during different phases of the cell cycle[24]. As a cell needs to complete the cell cycle in order to grow and divide, it is unsurprising that many of these genes are also highly correlated to cell proliferation. Indeed, a strong correlation was found when comparing the CCS signature to a proliferation metagene in the pan-cancer samples from this study (Pearson correlation = 0.95, $P < 0.001$, data not shown). This makes it difficult to disentangle specific cell cycle activity from cell proliferation, however, in the case of tumour cells with optimal growth conditions, the CCS signature likely gives an accurate measure of both simultaneously. We base this supposition on fundamental cancer cell cycle biology, where sustained proliferative signalling is a hallmark and to achieve this goal tumour cells mutate keys steps of the cell cycle to prevent the cell from exiting the cycle and entering a quiescent or senescent state. Specifically, continuous cell division in cancer cells is accomplished by (i) circumventing the DNA damage checkpoint and (ii) promoting S-phase entry (for review see here ref. [1]). The former is largely achieved via mutations of p53 and it's associated pathways[25,26], while the latter is similarly achieved through genomic alterations that induce E2F-dependent transcription[4]. The resulting continuously cycling state and rate at which it occurs is likely reflected well in cell cycle gene expression and by extension, genomic or immunohistochemical markers of proliferation such as Ki67. One limitation to this interpretation, though is that the CCS signature is comprised of so many genes that tumours where mitogens impact the later stages of the cycle, such as in G2[27], could have a CCS which does not reflect the overall proliferative or cell cycle activity of the cell as a smaller number of genes will be altered. In addition, all cells are not cycling at the same rate throughout the tumour, and our method gives a readout of activity across the entire sample that could be influenced by pockets of highly cycling cells. In short, the most appropriate way to interpret the CCS signature in this study is as a marker of proliferation which is correlated to cell cycle activity in cancer cells, but we acknowledge that this is an oversimplification, and further studies are required to understand how the score more directly relates to cell cycle activity.

The limitations of our study are as follows: First, the samples in this study are not matched tumour and normal tissue from the same patient, which means that we need a large sample size within each normal tissue type to derive the most accurate median values. Whilst this is true for many of the tissue types, cervical tissue ($N = 13$) stands out as one where caution should be taken with over-interpretation of the results. Here, the median value could change as more normal cervical samples are added to the cohort, and the median value we used is unlikely to be an accurate true reflection of the variation in normal cervical tissue. Related to this, our choice of using the median value also assumes that normal tissue is comprised a single cell population where cell cycle activity is uniform across the sample. This is unlikely to be the case and does not take into account that subgroups of normal cells with different genomic or phenotypic characteristics may be present in the sample and cycling at different rates. It is equally important to note that hormonally influenced cell types from e.g. gynaecological tissues, could also display differences in cell cycle activity on the basis or pre- or post-menopausal status, we have not taken this in account. Second, we study broad chromosomal gains and losses rather than gene-centric copy number changes—this comes from our previous experience with this data and wanting to avoid a situation where the most changed genes in Group 2 would all come from the same chromosomal location. Third, as noted above, we are focusing on cell cycle activity changes by applying the CCS signature to mRNA data extracted from an entire tumour sample as opposed to single cells. This means we get an average signal across all tumour cells and that the variance that would be seen in cell cycle activity at a single-cell level is not taken into account. The main strengths are: First, we apply a methodology that broadly reclassifies tumour cell cycle activity in terms of its tissue of origin in order to derive basic biological insight. Second, we use multiple 'omics data types to assess and further understand the differences between tumours of low and high BC-CCS; and third, we describe a new hypothesis of cell cycle activity having a ceiling that tumours may be unable to push past.

In summary, this study describes the reclassification of tumour cell cycle activity in terms of its normal tissue of origin. We show that, in general, gynaecological cancers show the largest change in this activity and that it is likely driven by sex hormones, chromosomal arm-level alterations, and individual gene expression differences. Finally, we propose a new hypothesis of there being an upper-limit or "ceiling" on cell cycle activity in tumours at a pan-cancer level.

METHODS

Study population and specimens

The UCSC Toil RNA-seq Recompute Compendium is a collection of study of origin batch effect-corrected RNA-seq samples from three datasets including The Cancer Genome Atlas (TCGA) ($N = 10535$), Therapeutically Applicable Research To Generate Effective Treatments (TARGET) ($N = 734$) and Genotype-Tissue Expression (GTex) ($N = 7862$)[8]. The compendium contains tumours from 24 different cancer types, including Adrenocortical carcinoma (ACC), Bladder Urothelial Carcinoma (BLCA), Brain lower grade Glioma (LGG), Breast invasive carcinoma (BRCA), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Colon adenocarcinoma (COAD), Oesophageal carcinoma (ESCA), Glioblatoma multiforme (GBM), Head and Neck squamous cell carcinoma (HNSC), Kidney Chromophobe (KICH), Kidney renal clear cell carcinoma (KIRC), Kidney renal papillary cell carcinoma (KIRP), Liver hepatocellular carcinoma (LICH), Lung adenocarcinoma (LUAD), Lung squamous cell carcinoma (LUSC), Ovarian serous cystadenocarcinoma (OV), Pancreatic adenocarcinoma (PAAD), Prostate adenocarinoma (PRAD), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), Testicular Germ Cell tumours (TGCT), Thyroid carcinoma (THCA), Uterine Carcinosarcoma (UCS), and Uterine Corpus Endometrial Carcinoma (UCEC) as well as normal tissue samples. All data are publicly available[28] and the quality control, normalisation and gene level counts were performed by the Toil investigators as described in the original publication[8]. Note that for some cancer types there are only GTEx normals or TCGA normals, all cancer types do not have normal samples from both studies.

Cell cycle score (CCS) and baseline corrected-cell cycle score (BC-CCS)

The CCS signature, along with its gene composition and method of application, has been previously extensively described[7,9,10]. Briefly, the signature is comprised of 463 cell cycle-related genes that were originally identified through the aggregation of three different pathway-related databases - Cyclebase 3.0, Kyoto Encyclopaedia of Genes and Genomes (KEGG) and HUGO gene nomenclature committee (HGNC)[24,29,30]. Many of these genes are proliferation-related and accordingly when comparing the CCS signature to a proliferation metagene in the pan-cancer samples from this study, a strong correlation was found (Pearson correlation = 0.95, $P < 0.001$, data not shown). The relationship between cell cycle activity and cell proliferation is further described in the discussion, however, we note that the simplest interpretation of the signature is as a marker of proliferation which is correlated to cell cycle activity in cancer cells. The 446 of the 463 original CCS signature genes were present in this study (Supplementary Data 1) and the final signature score was derived by summing up expression values of the signature genes resulting in a single CCS value on a per tissue sample/tumour basis. As these samples have already been normalised and standardised together as part of the Toil pipeline to make them directly comparable, the Baseline Corrected-Cell Cycle Score (BC-CCS) was calculated for each tumour sample by subtracting the median CCS of its GTEx normal tissue from the CCS of the tumour sample. For example, *BC-CCS Bladder Tumour 1 = CCS Bladder Tumour 1 − median CCS all normal Bladder Tissue*. The BC-CCS was additionally adjusted for tumour purity to account for differences in tumour epithelial content using the ABSOLUTE algorithm values previously derived by the pan-cancer investigators[31] and average purity values on a per-cancer basis are shown in Supplementary Table 2. This adjustment was performed by multiplying the BC-CCS for each tumour by it's individual purity value. Finally, both the CCS and BC-CCS continuous variables were scaled to values between 0 and 1 to aid with plotting and data visualisation.

Study of origin batch effect and outlier assessment

Principal component analysis (PCA) was performed with the *gmodels* R-package using (i) the genes with the highest variation in the dataset ($N = 6364$) and (ii) the genes included in the Cell Cycle Score (CCS, $N = 446$)[9]. Study of origin batch effects and the presence of outliers were assessed through manual examination of the PCA plots and overlaying study origin (GTEx, TCGA-normal, TCGA-cancer), tissue type (each of the 19 tissue types included in the study) or baseline corrected change in Cell Cycle Score data. Uniform Manifold Approximation and Projection (UMAP) was also applied to the dataset using the *umap* R-package as it provides high accuracy in separating the features of a complex data while identifying batch effects[32].

Mutation and chromosomal arm-level analyses

We used publicly available fully processed[12] mutation and chromosomal arm-level alteration data from the TCGA pan-cancer samples housed in the Genome Data Commons (GDC) database (https://gdc.cancer.gov). Briefly, arm-level alterations were defined by the original authors[33] as clustered somatic copy number alterations (SCNAs) where if the mean SCNA length was >80% of the chromosome, it was considered a positive for an arm alteration and <20% was considered negative for an arm alteration. Amplifications and deletions were considered separately for each arm using this methodology. Arms were then denoted as −1 if lost, +1 if gained, and 0 if negative for an alteration (non-aneuploid). These data were used to compare the DNA-level differences between the two groups with the lowest and highest BC-CCS, adjusting for the number of individual tumours with each cancer type of these two groups (Group 1: HNSC, KICH, KIRP UCEC and Group 2: CESC, OV, UCS) and for multiple testing using false discovery rate (FDR). Of note, for the mutational analysis, we focused on 299 cancer driver genes, that were manually annotated by experts in the field[12].

Gene expression and Gene Set Enrichment analyses (GSEA)

Differential gene expression analysis was performed using the *limma*[34] R-package in order to understand the mRNA differences between the two groups with the lowest and highest BC-CCS. The model matrix included additional variables to adjust for individual cancer types (HNSC, KICH, KIRP UCEC, CESC, OV, and UCS), and the oestrogen receptor gene *ESR1* as a continuous variable in order to adjust for the general impact of sex hormones on the cell cycle. Results were corrected for multiple testing and genes with Log_2 fold-change (FC) > 2 and FDR < 5% were considered significant. Note that we applied this strict 5% FDR in order to derive a list of genes we could be more confident were differentially expressed. GSEA was used to evaluate the enrichment of cancer hallmark pathways within the differentially expressed genes using the *fgsea*[35] R-package. Pathway rankings were ordered based on FC and p values. Results were corrected for multiple testing, and again we denote a stringent FDR < 10% threshold as statistical significance (GSEA uses 25% FDR as standard[36]).

Statistical analysis

Student's T-test was used to assess differences between normal and tumour continuous CCS, and Fisher's-exact test was applied to determine if gene mutations or chromosomal arm-level alterations were significantly different between the two BC-CCS groups. Genes found to be differentially expressed at the mRNA level between the same groups were tested for their correlation to the BC-CCS using Spearman's rank correlation test. To test for similarities between the CCS and a proliferation metagene a Pearson correlation test was performed (data not shown). All tests were 2-sided and $p < 0.05$ was considered as statistically significant. The data fulfilled the preconditions/assumption of the above tests. All measurements were taken from distinct individual samples. Boxplots should be interpreted as follows: horizontal lines denote median values; boxes extend from the 25th to the 75thpercentile of each group's distribution of values; vertical extending lines denote adjacent values (the most extreme values within 1.5 interquartile range of the 25th and 75th percentile of each group). Continuous CCS was normally distributed with low variation between groups. All statistical analyses were performed using R statistical software version 4.1.1[37].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

DATA AVAILABILITY

The data used in this study are publicly available on National Institute of Health (NIH) (Aneuploidy score and arm calls: https://gdc.cancer.gov/about-data/publications/pancan-aneuploidy - PANCAN_ArmCallsAndAneuploidyScore_092817.txt; Tumour purity data: https://gdc.cancer.gov/about-data/publications/pancanatlas) and UCSC Xena website (https://toil.xenahubs.net:443). No unique accession codes or identifiers are required to obtain the datasets.

CODE AVAILABILITY

Custom R-code to reproduce the results of this study is available using R statistical software version 4.1.1, are publicly available with unrestricted access on GitHub: https://github.com/arianlundberg/BC-CCS.

Received: 17 February 2022; Accepted: 13 July 2022;
Published online: 20 August 2022

REFERENCES

1. Matthews, H. K., Bertoli, C. & de Bruin, R. A. M. Cell cycle control in cancer. *Nat. Rev. Mol. Cell Biol.* https://doi.org/10.1038/s41580-021-00404-3 (2021).
2. Coller, H. A., Sang, L. & Roberts, J. M. A new description of cellular quiescence. *PLoS Biol.* **4**, e83 (2006).
3. Hanahan, D. & Weinberg, R. A. Hallmarks of cancer: The next generation. *Cell* **144**, 646–674 (2011).
4. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).
5. Lundberg, A. et al. The long-term prognostic and predictive capacity of cyclin D1 gene amplification in 2305 breast tumours. *Breast Cancer Res.* **21**, 34 (2019).
6. Helsten, T. et al. Cell-cycle gene alterations in 4864 tumors analyzed by next-generation sequencing: Implications for targeted therapeutics. *Mol. Cancer Ther.* **15**, 1682–1690 (2016).
7. Lundberg, A. et al. A pan-cancer analysis of the frequency of DNA alterations across cell cycle activity levels. *Oncogene* **39**, 5430–5440 (2020).
8. Vivian, J. et al. Toil enables reproducible, open source, big biomedical data analyses. *Nat. Biotechnol.* **35**, 314–316 (2017).
9. Lundberg, A. et al. Gene expression signatures and immunohistochemical subtypes add prognostic value to each other in breast cancer cohorts. *Clin. Cancer Res.* **23**, 7512–7520 (2017).
10. Tobin, N. P. et al. PAM50 provides prognostic information when applied to the lymph node metastases of advanced breast cancer patients. *Clin. Cancer Res.* **23**, 7225–7231 (2017).
11. Carter, S. L. et al. Absolute quantification of somatic DNA alterations in human cancer. *Nat. Biotechnol.* **30**, 413–421 (2012).
12. Bailey, M. H. et al. Comprehensive characterization of cancer driver genes and mutations. *Cell* **173**, 371–385.e18 (2018).
13. Berger, A. C. et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33**, 690–705.e9 (2018).
14. Knudsen, E. S. et al. Pan-cancer molecular analysis of the RB tumor suppressor pathway. *Commun. Biol.* **3**, 158 (2020).
15. Shen, S. & Clairambault, J. Cell plasticity in cancer cell populations. *F1000Res* **9**, F1000 Faculty Rev-635 (2020).
16. Aiello, N. M. et al. EMT subtype influences epithelial plasticity and mode of cell migration. *Dev. Cell* **45**, 681–695.e4 (2018).
17. Mu, P. et al. SOX2 promotes lineage plasticity and antiandrogen resistance in TP53- and RB1-deficient prostate cancer. *Science* **355**, 84–88 (2017).
18. Zajac, O. et al. Tumour spheres with inverted polarity drive the formation of peritoneal metastases in patients with hypermethylated colorectal carcinomas. *Nat. Cell Biol.* **20**, 296–306 (2018).
19. Malta, T. M. et al. Machine learning identifies stemness features associated with oncogenic dedifferentiation. *Cell* **173**, 338–354.e15 (2018).
20. Zhang, J. et al. The BRCA2-MEILB2-BRME1 complex governs meiotic recombination and impairs the mitotic BRCA2-RAD51 function in cancer cells. *Nat. Commun.* **11**, 2055 (2020).
21. Lightfoot, M., Montemorano, L. & Bixel, K. PARP inhibitors in gynecologic cancers: What is the next big development? *Curr. Oncol. Rep.* **22**, 29 (2020).
22. De Mello, R. A. et al. What will we expect from novel therapies to esophageal and gastric malignancies? *Am. Soc. Clin. Oncol. Educ. Book* **38**, 249–261 (2018).
23. Zhang, S. et al. BKM120 sensitizes glioblastoma to the PARP inhibitor rucaparib by suppressing homologous recombination repair. *Cell Death Dis.* **12**, 546 (2021).
24. Santos, A., Wernersson, R. & Jensen, L. J. Cyclebase 3.0: A multi-organism database on cell-cycle regulation and phenotypes. *Nucleic Acids Res.* **43**, D1140–D1144 (2015).
25. Zilfou, J. T. & Lowe, S. W. Tumor suppressive functions of p53. *Cold Spring Harb. Perspect. Biol.* **1**, a001883 (2009).
26. Hafner, A., Bulyk, M. L., Jambhekar, A. & Lahav, G. The multiple mechanisms that regulate p53 activity and cell fate. *Nat. Rev. Mol. Cell Biol.* **20**, 199–210 (2019).
27. Hitomi, M. & Stacey, D. W. Cellular ras and cyclin D1 are required during different cell cycle periods in cycling NIH 3T3 cells. *Mol. Cell Biol.* **19**, 4623–4632 (1999).
28. Goldman, M. J. et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020).
29. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res.* **44**, D457–D462 (2016).
30. Gray, K. A., Yates, B., Seal, R. L., Wright, M. W. & Bruford, E. A. Genenames.org: The HGNC resources in 2015. *Nucleic Acids Res.* **43**, D1079–D1085 (2015).
31. Aran, D., Sirota, M. & Butte, A. J. Systematic pan-cancer analysis of tumour purity. *Nat. Commun.* **6**, 8971 (2015).
32. Yang, Y. et al. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data. *Cell Rep.* **36**, 109442 (2021).
33. Taylor, A. M. et al. Genomic and functional approaches to understanding cancer aneuploidy. *Cancer Cell* **33**, 676–689.e3 (2018).
34. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
35. Korotkevich, G. et al. *Fast gene set enrichment analysis*. http://biorxiv.org/lookup/doi/10.1101/060012 (2016).
36. Subramanian, A. et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
37. R Development Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. http://www.R-project.org (2008).

ACKNOWLEDGEMENTS

This work was supported by the Iris, Stig och Gerry Castenbäcks Stiftelse for cancer research (N.P.T.); the King Gustaf V Jubilee Foundation (N.P.T.); the Stockholm Cancer Society (Cancerföreningen i Stockholm to L.S.L.); the Swedish Cancer Society (Cancerfonden, N.P.T. grant number: 200802; L.S.L. grant number: 190140); the Swedish Research Council (Vetenskapsrådet, grant number 2020-02466 to L.S.L); the Swedish Research Council for Health, Working life and Welfare, (FORTE, grant number 2019-00477 to L.S.L.); ALF medicine (grant number LS2018-1157 to L.S.L.) and the Gösta Milton Donation Fund (Stiftelsen Gösta Miltons donationsfond, to L.S.L.

AUTHOR CONTRIBUTIONS

A.L., J.J.J.Y., L.S.L., and N.P.T. contributed the study concept and design. A.L., J.J.J.Y., and N.P.T. contributed to the acquisition and analysis of data. All authors interpreted the data, drafted the manuscript, read and approved the final version.

FUNDING

Open access funding provided by Karolinska Institute.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41698-022-00302-7.

Correspondence and requests for materials should be addressed to Nicholas P. Tobin.

Reprints and permission information is available at http://www.nature.com/reprints

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