A pan-cancer analysis of the frequency of DNA alterations across cell cycle activity levels

Arian Lundberg[1,2] · Linda S. Lindström[3] · Joel S. Parker[4] · Elinor Löverli[1] · Charles M. Perou[4,5] · Jonas Bergh[1,6] · Nicholas P. Tobin[1]

Received: 16 April 2020 / Revised: 4 June 2020 / Accepted: 10 June 2020 / Published online: 24 June 2020

Abstract

Pan-cancer genomic analyses based on the magnitude of pathway activity are currently lacking. Focusing on the cell cycle, we examined the DNA mutations and chromosome arm-level aneuploidy within tumours with low, intermediate and high cell-cycle activity in 9515 pan-cancer patients with 32 different tumour types. Boxplots showed that cell-cycle activity varied broadly across and within all cancers. *TP53* and *PIK3CA* mutations were common in all cell cycle score (CCS) tertiles but with increasing frequency as cell-cycle activity levels increased ($P < 0.001$). Mutations in *BRAF* and gains in 16p were less frequent in CCS High tumours ($P < 0.001$). In Kaplan–Meier analysis, patients whose tumours were CCS Low had a longer Progression Free Interval (PFI) relative to Intermediate or High ($P < 0.001$) and this significance remained in multivariable analysis (CCS Intermediate: HR = 1.37; 95% CI 1.17–1.60, CCS High: 1.54; 1.29–1.84, CCS Low = Ref). These results demonstrate that whilst similar DNA alterations can be found at all cell-cycle activity levels, some notable exceptions exist. Moreover, independent prognostic information can be derived on a pan-cancer level from a simple measure of cell-cycle activity.

Introduction

The Nobel prize winning research of Hartwell [1], Nurse [2, 3] and Hunt [4] in the nineteen seventies and eighties fundamentally changed our understanding of the cell cycle and provided broad insight into the molecules governing its regulation. These seminal discoveries have shaped our modern view of the cell cycle and its separation into four distinct phases commonly referred to as G1, S, G2 and M. Transitions between these phases are governed by the cyclin family of proteins along with their binding partners the cyclin dependent kinases (CDKs) [5]. Disruptions to the function of cyclin-CDK holoenzymes or other cell cycle pathway members can lead to impaired control over the cycle and sustained proliferation—a hallmark of cancer [6].

Large scale pan-cancer studies have sought to understand human malignancies at a molecular level through the integration of multiple high-throughput data types. This approach has yielded a number of clinically relevant findings including the coalescence of lung squamous, head and neck, and some bladder cancers into a single pan-cancer subtype and the ability to classify tumours into prognostic subgroups at a pan-cancer level [7]. More recently, data from over 11,000 patients has shown actionable mutations in up to 57% of tumours [8], a positive correlation between aneuploidy and cell-cycle genes [9], and frequent co-alterations in the p53 and cell-cycle pathways [10]. To date, the analysis of genomic aberrations in these studies have typically focused on all pan-cancer tumours at once [8],

Supplementary information The online version of this article (https://doi.org/10.1038/s41388-020-1367-4) contains supplementary material, which is available to authorized users.

✉ Nicholas P. Tobin
nick.tobin@ki.se

[1] Department of Oncology and Pathology, Karolinska Institutet and University Hospital, Stockholm, Sweden

[2] Department of Radiation Oncology, Stanford School of Medicine, Stanford, CA, USA

[3] Department of Biosciences and Nutrition, Karolinska Institutet and University Hospital, Stockholm, Sweden

[4] Department of Genetics, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[5] Department of Pathology and Laboratory Medicine, Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[6] Department of Public Health, Oxford University, Oxford, UK

Table 1 Clinical characteristics of all patients split by CCS—*pan-cancer*.

Variables	Pan-cancer (n = 9515)			
	Low n (%)	Intermediate n (%)	High n (%)	
	3145 (33)	3184 (33.5)	3186 (33.5)	p
Age				
≤54	1290 (41)	876 (28)	1061 (34)	**<0.001**
54–66	1044 (33)	1062 (33)	996 (31)	
≥66	808 (26)	1236 (39)	1119 (35)	
Missing cases = 23				
Gender				
Male	1771 (56)	1372 (43)	1494 (47)	**<0.001**
Female	1374 (44)	1812 (57)	1692 (53)	
Pathological stage				
Stage I	859 (45)	601 (26)	444 (20)	**<0.001**
Stage II	480 (25)	820 (35)	768 (35)	
Stage III	419 (22)	639 (28)	575 (27)	
Stage IV	150 (8)	260 (11)	382 (18)	
Missing cases & excluded cases[a] = 3118				
Radiotherapy				
No	1954 (73)	2047 (73)	1820 (65)	**<0.001**
Yes	709 (27)	770 (27)	993 (35)	
Missing cases = 1222				

[a] I/II NOS-Stage 0/IS/X, in bold significant $p < 0.05$.

within subgroups of tumours that have clustered together on the basis of DNA, RNA and protein expression—termed the iClusters [8], or within tumours with a common genetic alteration such as chromosome 3p loss [9]. Given the varying degrees of oncogenic pathway activation/suppression across cancer types [10], we hypothesised that basing genomic analyses on the magnitude of pathway activity may also provide important biological information and clinical insight. In view of the fundamental biological role of the cell cycle in cancer and the frequent genomic alterations of its pathway members, it represents a compelling choice for a pathway activity-based analysis.

Here, in order to test our hypothesis, we compare the most prevalent genomic alterations in tumours with low, intermediate and high levels of cell-cycle activity by integrating data from multiple genomic platforms in over 9000 tumours from The Cancer Genome Atlas (TCGA). Specifically, we examine gene expression levels, gene mutational frequency and chromosome arm-level alterations across pan-cancer tumours grouped into low, intermediate and high tertiles of cell-cycle activity on the basis of our cell cycle score (CCS) gene signature [11, 12]. Finally, we also determine the clinical relevance of this signature across and within cancer types using survival analyses including Kaplan–Meier graphs and multivariable Cox proportional hazards modelling adjusting for patient and tumour characteristics.

Results

Cohort clinico-pathological characteristics in relation to CCS subgroups

In line with our aim to compare genomic alterations in tumours with differing levels of cell-cycle activity we applied our CCS signature (genes are shown in Supplementary Table 1) to gene expression data from the tumours of 9515 pan-cancer patients. Clinico-pathological characteristics for the pan-cancer cohort split by low, intermediate and high CCS tertile classifications are shown in Table 1 and a CONSORT diagram showing the exclusion criteria for this study is shown in Supplementary Fig. 1. Statistically significant associations were found between patient age, gender, pathological stage, radiotherapy and CCS subgroups (Table 1, Chi-squared test: $P < 0.001$ for all comparisons). After adjusting for cancer type, only stage and radiotherapy remained statistically significant whereby CCS High tumours were more likely to be stage IV and to have received radiotherapy (data not shown).

Broad variation in cell-cycle activity across cancers and COCA subtypes

We next assessed tumour cell-cycle activity by creating pan-cancer, Cluster of cluster assignment (COCA) and iCluster boxplots using the continuous CCS. We found the highest levels of cell-cycle activity in Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC), Testicular Germ Cell tumours (TGCT), Head and Neck squamous cell carcinoma (HNSC) and Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC) tumours and the lowest in Kidney Chromophobe (KICH), Pheochromocytoma and Paraganglioma (PCPG), Kidney renal papillary cell carcinoma (KIRP) and Prostate adenocarcinoma (PRAD) tumours (Fig. 1a). Similar results were found using the COCA algorithm—a classification strategy that clusters samples by integrating information from multiple individual cross platform technologies, with CA17 (TCGT) and CA4 (PAN-SCC, mainly HNSC, LUSC and CESC tumours) forming the top two subgroups with the highest cell-cycle activity (Fig. 1b). CA10 (Breast Invasive Carcinoma (BRCA), basal-like) and CA25 (Haematologic/lymphatic, mainly Thymoma (THYM) and DLBC tumours), also showed high cell-cycle activity, whilst CA1 (CNS/Endocrine, mainly PCPG tumours), CA14 (PRAD) and C21 (PAN-Kidney) showed the lowest levels of all COCA subtypes (Fig. 1b). Analogous results were noted using the

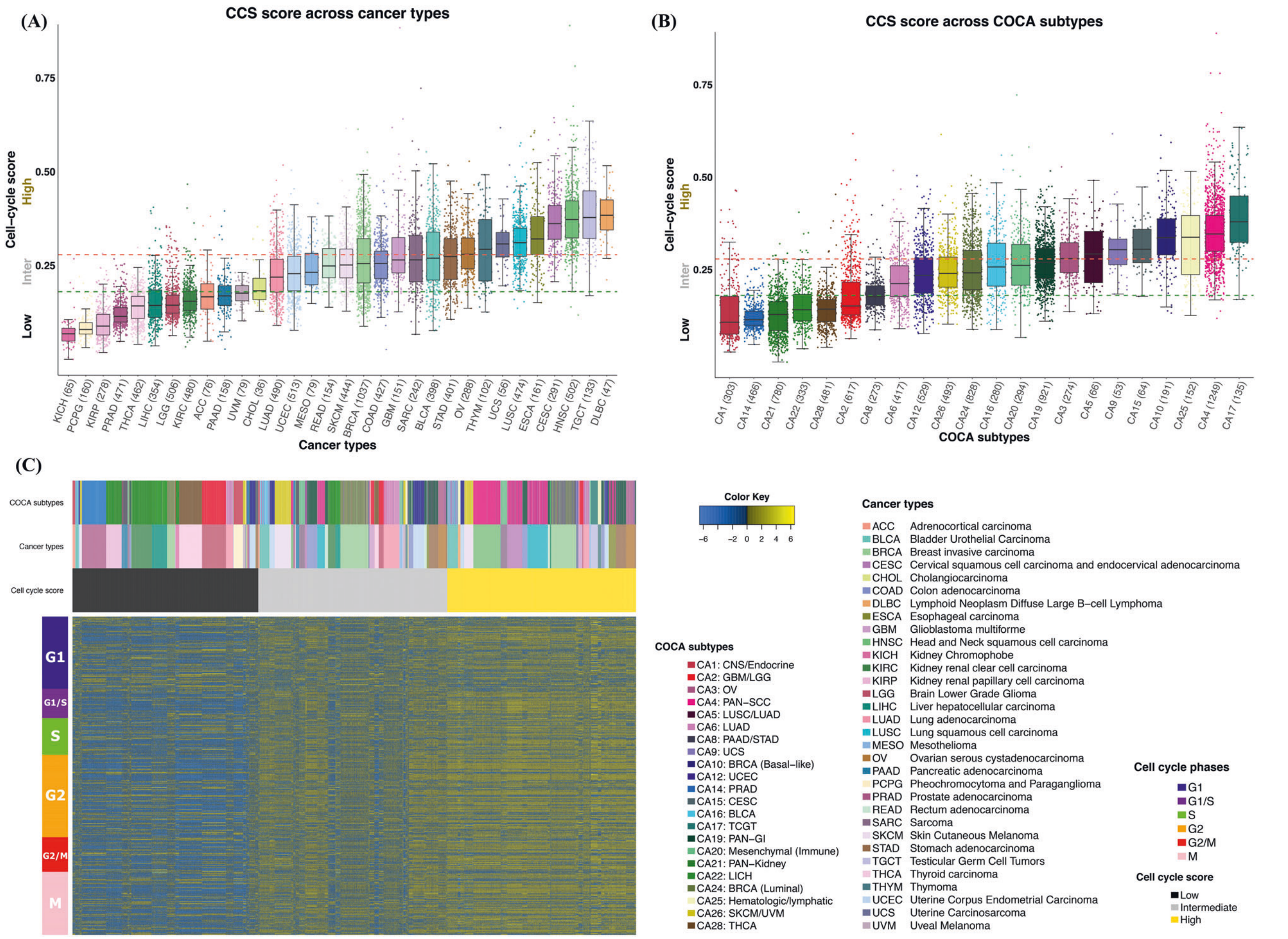

Fig. 1 CCS score across cancer types and COCA subtypes. Boxplots comparing CCS across (**a**) pan-cancer types and (**b**) COCA subtypes. Numbers in parentheses represent number of tumours in each cancer type and/or COCA subtype. **c** Heatmap of CCS genes across pan-cancer tumours. Heatmap colside colours (horizontal, above heatmap) represent cell cycle score, cancer types and COCA subtypes as indicated in figure legend. Rowside colours (vertical, left hand side of heatmap) represent cell cycle phases.

iCluster classification strategy (Supplementary Fig. 2). Examining cell-cycle activity clusters using heatmap analysis demonstrated that tumours with low levels of cell-cycle activity (and thus classified as CCS Low) show low expression of the majority of genes in all cell cycle phases (G1–M), whilst the opposite is true for tumours with high levels of cell-cycle activity (Fig. 1c, compare tumours with black column-side colour to those with yellow).

TP53 and *PIK3CA* mutations display increasing frequency across cell-cycle activity subgroups

To more clearly delineate the frequency of DNA mutations in relation to the magnitude of cell-cycle activity we next examined the mutational frequency of 299 well defined oncogene and tumour suppressor driver genes within CCS subgroups. *TP53* was found to be the most mutated gene in all three CCS subgroups and displayed an increase in mutational frequency with increasing CCS activity (Fig. 2a, Supplementary Table 2, Chi-squared test: $P < 0.001$). In CCS Low tumours 40% of *TP53* mutations were found in LGG, whereas in CCS High tumours *TP53* mutations were most common in HNSC (18%), LUSC (17%) and BRCA (13%) (Highlighted in Fig. 2a). *PIK3CA* was the second-most commonly mutated gene in CCS Intermediate and High tumours and fifth-most common in CCS Low tumours (Fig. 2a). It is also more frequently mutated in CCS Intermediate and High tumours relative to CCS Low (Supplementary Table 2, $P < 0.001$). *PIK3CA* mutations in BRCA and Uterine Corpus Endometrial Carcinoma (UCEC) were common across all CCS subgroups and were additionally found in HNSC and CESC in CCS High tumours (Fig. 2a). Of interest, whilst *BRAF* mutations were prominent in Low and Intermediate subgroups as the third and eleventh most mutated gene respectively, it was absent from the top 15 in CCS High tumours (Fig. 2a, red arrows, Supplementary Table 2, $P < 0.001$). These findings suggest genes other than *BRAF* are more commonly mutated in tumours with high

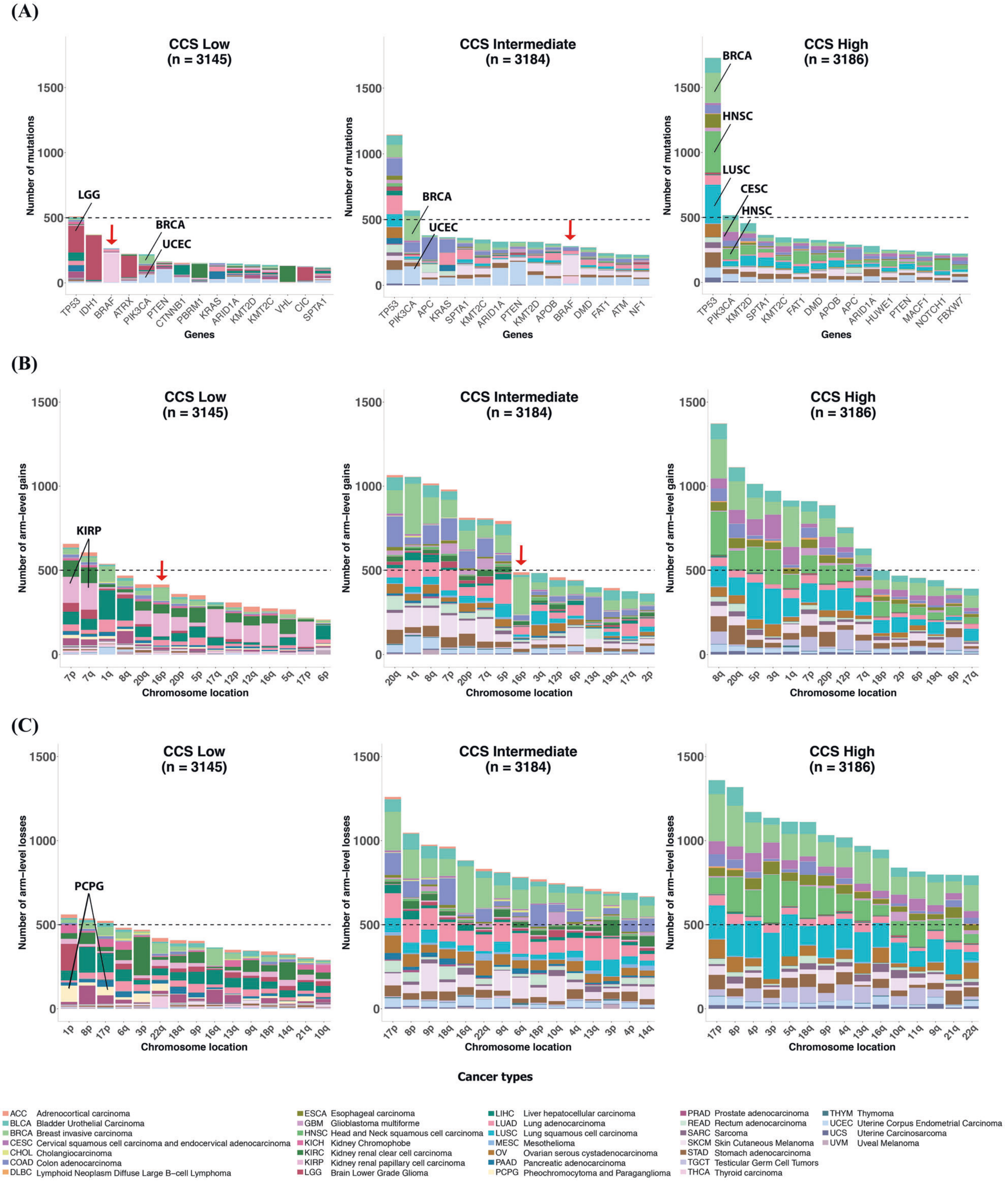

Fig. 2 Top 15 most commonly mutated genes or chromosomal arm-level alterations within CCS subgroups. Pan-cancer tumours were divided into tertiles on the basis of low, intermediate or high CCS. Within each subgroup the Top 15 (**a**) most frequently mutated oncogenes and tumour suppressor genes, (**b**) arm-level gains and (**c**) arm-level losses are shown. Cancer type colour key is shown at the bottom of the figure. Red arrows indicate *BRAF* mutations and 16p gains in CCS Low and Intermediate subgroups.

cell-cycle activity. Of note, the increased number of *BRAF* mutations in CCS Low tumours is mainly driven by a single tumour type—Thyroid carcinoma (THCA) (THCA, pink colour under red arrow in Fig. 2a), whereas in CCS Intermediate (and High) tumours BRAF mutations are mostly found in Skin Cutaneous Melanoma (SKCM)

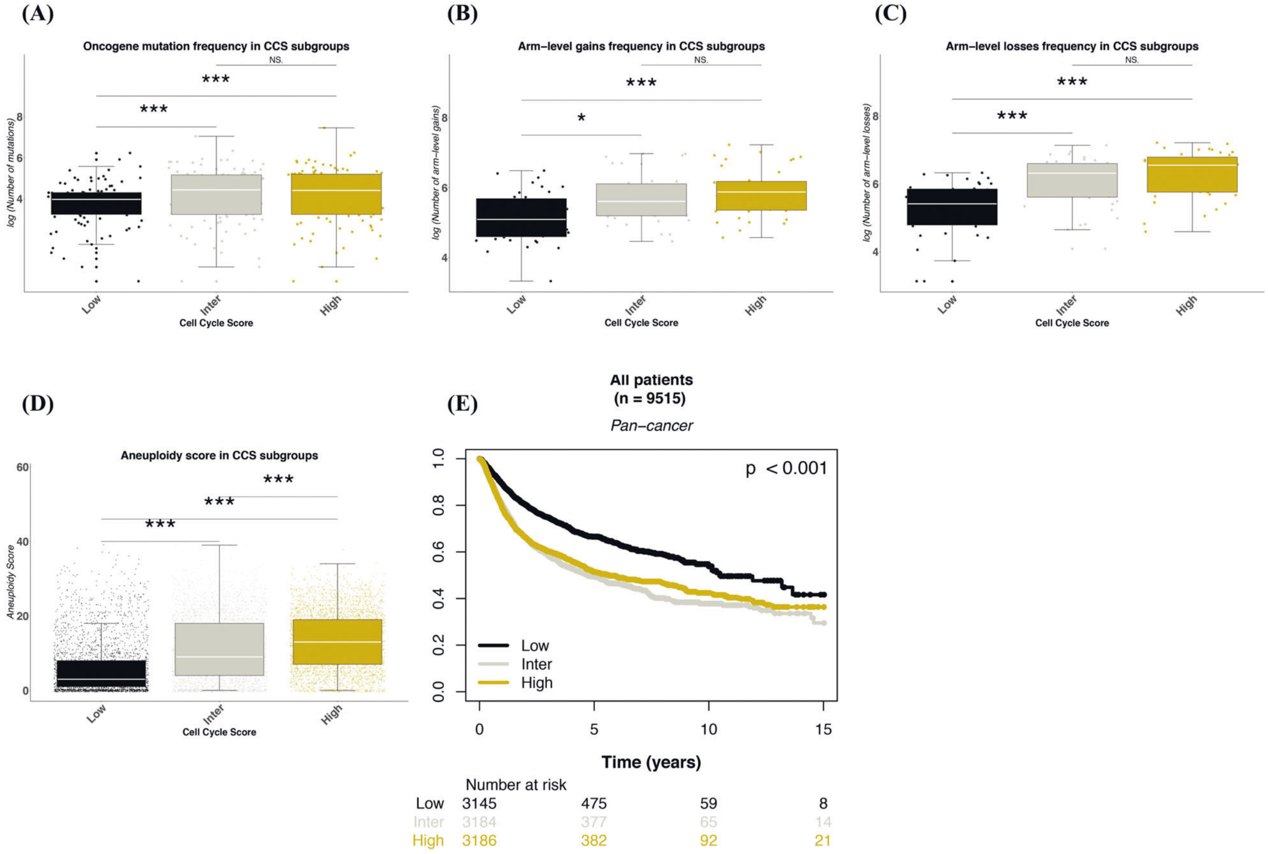

Fig. 3 Boxplots comparing frequency of DNA alterations across CCS subgroups. Pan-cancer tumours were divided into tertiles on the basis of low, intermediate or high CCS. Within each subgroup the number of (**a**) total mutations in the top 50 most mutated oncogenes or tumour suppressor genes, (**b**) total chromosomal arm-level gains, (**c**) total chromosomal arm-level losses and (**d**) aneuploidy score are shown. **e** Kaplan–Meier analysis of CCS subgroups with Progression free Interval (PFI) as clinical endpoint. Low/Inter/High = Low/Intermediate/High CCS subgroups, *p* values in boxplots (based on ANOVA with post-hoc Tukey HSD test) = NS > 0.05, * < 0.05, ** < 0.01, *** < 0.001; *p* value in the Kaplan–Meier curves refer to long-rank tests.

(Supplementary Table 3). The top 50 most frequently mutated genes in all CCS subgroups are shown in Supplementary Table 4.

Higher levels of chromosomal gains and losses in CCS Intermediate and High tumours

We next performed the same subgroup analysis, but this time focusing on chromosome arm-level gains and losses. All CCS subgroups showed a high number of gains to arms 20q, 8q and 7p and losses to arms 17p and 8p (Fig. 2b, c, respectively, all CCS subgroups). Moreover, these chromosomal aberrations all displayed an increase in frequency with increasing CCS activity (Supplementary Table 2, Chi-squared test: $P < 0.001$ for all comparisons, not adjusted for multiple testing). Overall, gains in KIRP (Fig. 2b, highlighted) and losses in PCPG cancers (Fig. 2c, highlighted) were more common in CCS Low tumours relative to CCS Intermediate and High subgroups, as could be anticipated given the low cell-cycle activity levels displayed by these tumour types and their grouping into the CCS Low tumour subgroup (Fig. 1a). Analogous to our *BRAF* mutation findings, gains to 16p (Fig. 2b, red arrows) were more common in CCS Low and Intermediate subgroups relative to the CCS High subgroup (Supplementary Table 2, $P < 0.001$). 29% of 16p gains are found in KIRP in CCS Low tumours, whereas they occur most commonly in BRCA in CCS Intermediate and High tumours (Supplementary Table 3). The frequency of chromosomal arm gains and losses in all CCS subgroups are shown in Supplementary Table 5.

Next, we examined genomic alterations more broadly within CCS subgroups and found the frequency of gene mutations and chromosomal arm gains and losses to be greater in CCS Intermediate and High groups relative to Low (Fig. 3a–c, Tukey HSD test, 3A top 50 DNA mutations: $P < 0.001$ and $P < 0.001$, 3B chromosomal gains: $P =$ 0.018 and $P < 0.001$ and losses 3C: $P < 0.001$ and $P < 0.001$ for Low vs. Intermediate and High, respectively). Similarly, using the recently derived aneuploidy score [9]—a measure of the total number of chromosome arms with arm-level copy number changes in a given sample, we also found a statistically significant increase with increasing CCS activity levels (Fig. 3d, $P < 0.001$ for all comparisons).

CCS signature provides independent prognostic information at pan-cancer level

We next assessed the relationship between CCS and Progression Free Interval (PFI) using Kaplan–Meier and multivariable Cox proportional hazard regression model analyses. In univariate Kaplan–Meier analysis patients whose tumours were classified as CCS Low had a significantly longer PFI relative to those classified as CCS Intermediate or High (Fig. 3e, log-rank test: $P < 0.001$). This significance remained when adjusting for tumour type, age, gender, pathological stage and radiotherapy in Cox proportional hazard analysis (Table 2, CCS Intermediate: HR 1.37 95% CI 1.17–1.60, CCS High: HR 1.54 95% CI 1.29–1.84, tumour type not shown). The upper age tertile (≥66) remained statistically significant in the same model (HR 1.19 95% CI 1.05–1.35 vs. Ref), as did all pathological stages vs. the Stage I model reference group. As many cancers contain additional molecular subgroups (e.g. breast cancer) we also performed a similar analysis but adjusting for COCA subtypes rather than pan-cancer types and found comparable independent prognostic capacity for the CCS (data not shown).

In order to more closely examine individual cancer types where the signature splits tumours into two or three CCS subgroups, we again performed Kaplan–Meier and Cox proportional hazard modelling but this time focusing on individual cancers. CCS provided significant independent prognostic information in four cancer types: Kidney renal clear cell carcinoma (KIRC) ($P = 0.042$), LGG ($P < 0.001$), Sarcoma (SARC) ($P = 0.001$) and Uveal Melanoma (UVM) ($P = 0.013$, Supplementary Fig. 3, alphabetical ordering, adjusted for multiple testing). Finally, as the CCS subgroups are based on a tertile split of cell-cycle activity on a pan-cancer level, we hypothesised that deriving subgroups in this manner may provide superior prognostic information to a simple tertile split within (intra) each cancer type. To test this hypothesis, we compared our pan-cancer CCS tertile subgroups to intra-cancer CCS tertile subgroups. We found that whilst both cut-offs provide significant prognostic information in the above four cancer types (Compare Kaplan–Meier curves for pan-cancer CCS to intra-cancer CCS, Supplementary Fig. 4), a pan-cancer cut-off provides more prognostic information calculated by the likelihood ratio (LR) test, in KIRC (LR = 24.7), LGG (LR = 31.1), SARC (LR = 18.5) and UVM cancers (LR = 17.1), Table 3, compare pan-cancer column to intra-cancer). These findings suggest that deriving transcriptional biomarker cut-points on a pan-cancer level may be advantageous relative to deriving them in a single cancer type. For the sake of completeness, hazard ratios and 95% confidence intervals for pan-cancer and intra-cancer tertile subgroups in individual cancers are shown in Supplementary Table 6.

Table 2 Multivariate evaluation of prognostic markers in patients characterised by Cell Cycle Score.

Variables	Pan-cancer ($n = 5421$)[a]			
	N (%)	HR	95% CI	*p*
Age				
≤54	1679 (31)	Ref	–	–
54–66	1761 (32)	1.04	0.91–1.18	0.551
≥66	1981 (37)	1.19	1.05–1.35	**0.008**
Missing cases = 23				
Gender				
Male	2757 (51)	Ref	–	–
Female	2664 (49)	0.96	0.87–1.07	0.483
Pathological stage				
Stage I	1561 (29)	Ref	–	–
Stage II	1852 (34)	1.60	1.38–1.86	**<0.001**
Stage III	1378 (25)	2.41	2.08–2.79	**<0.001**
Stage IV	630 (12)	5.04	4.21–6.03	**<0.001**
Missing cases = 3118				
Radiotherapy				
No	3997 (74)	Ref	–	–
Yes	1424 (26)	0.97	0.84–1.11	0.658
Missing cases = 1222				
Cell cycle score				
Low	1505 (28)	Ref	–	–
Intermediate	2013 (37)	1.37	1.17–1.60	**<0.001**
High	1903 (35)	1.54	1.29–1.84	**<0.001**

In bold significant $p < 0.05$.

Ref reference groups, *N* number of patients, *HR* hazard ratio, *CI* confidence interval.

[a] Adjusted for cancer types.

Discussion

The present study integrates gene expression, DNA mutation, DNA copy number and clinico-pathological data from 9515 pan-cancer patients in order to better understand the DNA level alterations present in tumours with low, intermediate and high cell-cycle activity. Our main findings show first, that cell-cycle activity varies broadly across and within cancer types; second, that *TP53*, *PIK3CA* and chromosomal alterations (including gains to 20q, 8q, 7p and losses to arms 17p and 8p) occur with increasing frequency in tumours with increasing cell-cycle activity; third, whilst in general similar mutations/arm-level alterations are present within tumours

Table 3 Prognostic value of Cell Cycle signature (CCS) based on likelihood ratio (LR-X^2) and concordance index (C-index).

Pan-cancer and *intra-cancer*								
Models	Cell cycle score							
	Total	Events	*Pan-Cancer*			*Intra-cancer*		
Univariate			C-index	LR-X^2	p	C-index	LR-X^2	p
KIRC	480	148	0.589	24.7	**<0.001**	0.567	9.9	**0.001**
LGG	506	192	0.609	31.1	**<0.001**	0.627	21.9	**<0.001**
SARC	242	124	0.610	18.5	**<0.001**	0.614	14.3	**<0.001**
UVM	79	24	0.732	17.1	**<0.001**	0.707	11.5	**<0.001**

Events: Progression Free Interval (PFI) events in which patients had a new tumour whether it was a progression of disease, local recurrence, distant metastasis, new primary tumours all sites, or died with the cancer without new tumour event. In bold significant $p < 0.05$.

KIRC kidney renal clear cell carcinoma, *LGG* brain lower grade glioma, *SARC* sarcoma, *UVM* uveal melanoma, *CCS* cell cycle score, *LR-X^2* likelihood ratio, *C-index* concordance index.

with low, intermediate and high cell-cycle activity, mutations in *BRAF* and gains in 16p were less frequent in tumours with high cell-cycle activity; and fourth, that deriving cut-points for biomarkers on a pan-cancer level may provide more prognostic information than deriving them within specific cancer types. These analyses are the first to provide broad insight on the genetic alterations occurring within tumours grouped on the basis of cell-cycle activity in order to advance our understanding of a pathway that is frequently dysregulated in human malignancies.

In pan-cancer analyses, *TP53, PIK3CA, KRAS, PTEN* and *ARID1A* genes have all been previously demonstrated to be mutated in over 15 different cancer types [8]. These genes also featured heavily in our mutational analysis with *TP53* and *PIK3CA* mutations showing the high- mutational frequency across CCS subgroups. This implies that mutations in these genes are found in tumours with a broad range of cell -cycle activity and are not just associated with highly cycling cancers, despite their very clear links to cell-cycle progression [13, 14]. Whilst we found the *ARID1A* gene to be mutated in all CCS subgroups, *BRAF* was notable for only being found in the top 15 of the CCS Low and Intermediate subgroups, implying that other genes are more commonly mutated in tumours with high cell-cycle activity, such as *TP53* and *PIK3CA*. This result is partially driven by the cancer types found in each of the CCS subgroups, e.g. *BRAF* mutations are predominantly found in THCA cancers in the CCS Low subgroup and SKCM in CCS Intermediate and High tumours. It is important to highlight that CCS tumour subgroups were derived on the basis of biological cell cycle pathway activity alone. Our aim was to provide map/characterise the DNA aberrations within tumours on the basis of pathway magnitude, as such, even if a specific aberration is enriched owing to a certain tumour type, it is still one characteristic of tumours with low levels of cell-cycle activity, albeit one associated with a specific cancer type.

It has recently been demonstrated that tumour aneuploidy is inversely correlated to immune signalling genes and positively correlated to cell cycle and pro-proliferation pathways [9]. Our findings are in line with these showing a step wise increase in aneuploidy score with increasing CCS activity levels. Related to this, whilst most of predominant chromosome arm-level alterations we observed overlapped with those from the pan-cancer publication [9], our within subgroup analysis yielded some novel findings. In particular, and analogous to our mutational results, we found that specific gains (16p) were present in the CCS Intermediate and High subgroups only (Fig. 2b, c, red arrows). This raises the possibility that this chromosomal alteration could potentially be used as novel clinical biomarkers for more indolent tumours in cancers of unknown primary origin.

We found that our CCS gene expression signature, which has been previously applied in a breast cancer setting [11, 12], provided independent prognostic information on a pan-cancer level. This signature was originally conceived as a simple biological measure of cell-cycle activity in response to the dependence of more established commercial gene expression signatures on multiple cell cycle/cell proliferation genes for their prognostic capacity [15]. The signature genes were chosen through the aggregation of three different biological pathway databases [16–18], meaning that it is not cancer-specific and can be applied to any tissue sample. In keeping with its descriptive nature, we have not attempted to maximise the signature's prognostic capacity through selection of genes that are the strongest predictors of the study's clinical endpoint—PFI. Despite this, the signature performed well in both Kaplan–Meier and multivariable analyses, likely owing to its ability to select for faster growing, more aggressive tumours. As we are applying the CCS to a pan-cancer cohort, the prognostic capacity of the signature should be viewed in the context of all cancers and in general, not necessarily within specific cancer types. For example, when we split the continuous CCS into tertiles of activity the majority of prostate cancers (PRAD) are classified into the CCS Low subgroup and thus as “good” prognosis based on our analyses. Conversely, glioblastoma cancers (GBM) were predominantly classified into CCS Intermediate and High subgroups and thus as “poor” prognosis. In line with this, the median time to a PFI event for PRAD patients in the pan-cancer cohort is 18.4 months, whereas for GBM it’s 6.1 months [19]. As such, the prognostic capacity for the CCS signature when applied to all tumours cannot be determined on the basis of its strength within in a single cancer type, but only when considered in the context of all cancers. Interestingly, however, some cancers were split into two or three different CCS subgroups and upon further examination of these cancer types we found that deriving CCS tertiles of activity on a pan-cancer level may provide more prognostic information than deriving them within a specific cancer type. This could be of utility in a clinical setting where a gene transcript is being used as a biomarker for treatment response, such as the recent example of cyclin E expression and Palbociclib efficacy in metastatic breast cancer patients [20]. In this instance it is conceivable that re-defining a cyclin E cut-point on the basis of pan-cancer expression levels of the gene may more clearly delineate which patients are likely to be resistant to the drug.

When applying a gene expression signature to any dataset a choice regarding the best cutoffs for sample subgrouping is typically inherent to the analysis. Here, we chose to divide the continuous CCS into three equal groups resulting in low, intermediate and high expression subgroups. This decision was largely based on both our experience with other gene expression signatures in the breast cancer field where three subgroups are common, such as for the 21-gene recurrence score [21] and the biology-based gene expression modules [22]. Moreover, given that the CCS is continuum of values (as shown in Fig. 1) without any clear bimodal distribution, it does not make sense to force a simple binary high/low grouping on the data. Instead, we opted for tertiles that reflect this continuum with high and low expression groups and the addition of a third intermediate subgroup to cover the range of samples transitioning from low to high expression. Another important point to consider is that we are applying the CCS signature to data extracted from an entire tumour and as such are getting an average gene signal across the entire sample. This means that heterogeneity in terms of the cellular composition of the tumour and in terms of expression of the CCS in different tumour regions is not taken into account. Many newer technologies including single-cell sequencing and spatial transcriptomics can be used to examine tumour heterogeneity at single-cell resolution [23], however, as this type of data is not currently available for the tumours of the pan-cancer cohort we cannot assess the intratumour variation of the CCS in this material. A second, more traditional way to take heterogeneity into account is through the examination of whole tumour sections under the microscope. Given that the CCS is interlinked with cell proliferation, which in turn is an important component of tumour grading (in the form of mitotic count), one could speculate as to the merits of grading in addition in to or in place of applying the CCS. Unfortunately, tumour grading information was missing for over 50% of the tumours included in this study meaning it was not included in multivariable analyses. More importantly, grading systems differ greatly between cancer types, for example the three-level Nottingham histologic grade is used in breast cancer whilst Gleason grading with up to five different groups is used in prostate cancer. This renders the application of grade at a pan-cancer level currently unfeasible and relatedly, we have previously demonstrated the propensity of the CCS and other gene expression signatures to outperform ocular assessment of the proliferation marker Ki67 on whole tumour sections [11, 24].

There are three main strengths to our study; first, we utilise a novel methodology to examine the DNA alterations in subgroups of tumours that is based on the magnitude of cell-cycle activity both across and within cancer types; second, our analysis provides an expansive, descriptive overview of the frequency of DNA mutations and chromosomal gains and losses in subgroups of low, intermediate and high cell- cycle activity; and third, we demonstrate the translational relevance of our work by relating our CCS signature to a clinical survival endpoint —PFI. The limitations are as follows; first, our analysis focuses on DNA and RNA technologies only, with no protein or methylation array data included; second, we chose to study broad chromosomal gains and losses rather than gene-centric copy number changes—this was to avoid a situation where the most changed genes within a given CCS subgroup would all come from the same chromosomal location; third, we did not adjust the CCS for every molecular subgroup within every cancer type in multivariable analysis—this is a general limitation of any pan-cancer study, we did however perform additional analyses adjusting for COCA subtypes which captures more molecular heterogeneity than adjusting for cancer types alone and found analogous results; and fourth no external validation was performed for the CCS signature, although we are not aware of any other pan-cancer dataset where it could be validated and more importantly, we are not currently proposing it for use in a clinical setting—rather as a general tool to examine the cell-cycle activity of a given tumour.

In summary, this study describes the DNA mutations and chromosomal alterations found in tumours with low, intermediate and high levels of cell-cycle activity and also demonstrates the ability of a simple cell cycle gene expression signature to provide independent prognostic information at a pan-cancer level.

Materials and methods

Study population and specimens

The Pan-Cancer Atlas (PanCanAtlas) project compared and contrasted genomic and cellular differences between tumour types profiled as part of TCGA. The project consists of 11,069 patients with primary tumours from 32 different cancer types, including Adrenocortical carcinoma (ACC), Bladder Urothelial Carcinoma (BLCA), Breast Invasive Carcinoma (BRCA), Brain lower grade Glioma (LGG), Cervical squamous cell carcinoma and endocervical adenocarcinoma (CESC), Cholangiocarcinoma (CHOL), Colon adenocarcinoma (COAD), Esophageal 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), Lymphoid Neoplasm Diffuse Large B-cell Lymphoma (DLBC), Mesothelioma (MESO), Ovarian serous cystadenocarcinoma (OV), Pancreatic adenocarcinoma (PAAD), Pheochromocytoma and Paraganglioma (PCPG), Prostate adenocarcinoma (PRAD), Rectum adenocarcinoma (READ), Sarcoma (SARC), Skin Cutaneous Melanoma (SKCM), Stomach adenocarcinoma (STAD), Testicular Germ Cell tumours (TGCT), Thymoma (THYM), Thyroid carcinoma (THCA), Uterine Carcinosarcoma (UCS), Uterine Corpus Endometrial Carcinoma (UCEC) and Uveal Melanoma (UVM).

From the original 11,069 patients, 9515 were included in our study and reasons for exclusion were missing or no matching gene expression data ($n = 795$), copy number data ($n = 498$) or clinico-pathological information ($n = 261$). A CONSORT diagram showing the exclusion criteria for this study is shown in Supplementary Fig. 1. This cohort was chosen owing to its large sample size ensuring sufficient power for the statistical testing being performed. All clinical, gene expression, mutation and chromosome arm-level data from the PanCanAtlas study were taken from the publicly available database of the National Institutes of Health (NIH) (https://gdc.cancer.gov/about-data/publications/pancanatlas).

mRNA data, clustering and the Cell Cycle Score (CCS)

Fully processed, batch corrected, RNA-sequencing data were accessed from NIH genomic data commons (GDC) database (https://gdc.cancer.gov). All data quality control, normalisation and gene level counts were performed by the PanCanAtlas investigators as described in the their original publication [25]. iCluster were also retrieved from the same publication. COCA classifications were performed by the pan-can investigators as described in Hoadley et al. [7], resulting in 32 different tumour clusters. Clusters with <20 tumours were excluded from further analysis.

The CCS signature is comprised of 463 genes that were originally identified through the aggregation of three different pathway-related databases—KEGG, HGNC and Cyclebase 3.0 [16–18]. As these databases aim to describe general biology rather than being cancer focused, the CCS genes can be seen as representative of general cell-cycle activity and could be applied to any tissue sample (normal or tumour tissue). Whilst the signature has previously been applied in a breast cancer setting, the gene list has not been reduced or altered on the basis of those studies. For the sake of clarity and reproducibility all 463 signature genes are shown in Supplementary Table 1 along with annotations of which genes were present in previous breast cancer studies as well as the current pan-cancer dataset. Four hundred and forty-one of the 463 original CCS signature genes were present in, and extracted from, the pan-cancer dataset. Expression values were summed on an individual tumour basis to derive a single score of cell-cycle activity for each sample. This continuous variable was further divided into tertiles in order to classify tumours as having Low, Intermediate or High levels of cell-cycle activity on a broad, pan-cancer level. Cancer types where the pan-cancer CCS demonstrated independent prognostic information in multivariable Cox proportional hazard models were also assessed using within (intra-) cancer CCS tertiles: KIRC, LGG, SARC and UVM.

Mutational analysis

Fully processed mutational data derived from exome sequencing was taken from GDC database in a mutation annotation format file (MAF) (https://gdc.cancer.gov). All data quality control, processing and mutation calling was performed by the PanCanAtlas investigators as described in the their original publication [8]. We limited our analysis to 299 cancer driver genes manually annotated by experts in the pan-cancer field [8]. The MAF*tools* package in the R-statistical environment was used for mutation count calculations within CCS subgroups. A gene was counted as mutated (1) or not (0) for each tumour regardless of the number of mutations within that gene.

Chromosomal arm-level alterations and aneuploidy score

Fully processed chromosome arm-level alteration data and tumour aneuploidy scores were accessed from GDC database (https://gdc.cancer.gov) and were derived from Affymetrix SNP 6.0 arrays. All data quality control and processing was performed by the PanCanAtlas investigators as described in the original publication [26]. Chromosome arm-level alterations are presented as estimated ploidy values of +1, 0 and −1 for gains, non-aneuploidy and losses, respectively [9].

Statistical analysis

To assess differences among clinico-pathological characteristics of tumour samples and CCS subgroups χ^2 tests were employed. Clinical and survival data were retrieved from the GDC database (https://gdc.cancer.gov/about-data/publications/pancanatlas). Univariate Kaplan–Meier analysis was performed for the CCS in all pan-cancer tumours together and in individual cancer types with PFI censored at 15 years as the clinical endpoint, as previously recommended [19]. PFI is defined as the period during or after the course of a treatment given to patients in which the disease does not show any progression until a loco-regional recurrence and/or second malignancy occurs, or the patients die from any cause. Multivariable Cox proportional hazard models were used to determine the independent prognostic capacity of the CCS subgroups in all pan-cancer tumours together and in individual cancer types adjusting for cancer type, age (grouped in tertiles), gender, radiation therapy and pathological stage. Tumour grading information was missing for over 50% of pan-cancer samples and as such was not included in multivariable analyses. To compare the prognostic capacity of pan-cancer vs. intra-cancer CCS cutoffs we used the LR, which can be interpreted as a goodness-of-fit test. LR and concordance index (C-index) measures were extracted from the output of the coxph function of the *survival* package in R. Genomic alterations including the frequency of gene mutations and chromosomal arm gains and losses as well as aneuploidy score were compared between three CCS subgroups by using ANOVA with the post-hoc Tukey HSD test, all tests were two-sided and $p < 0.05$ was considered as statistically significant, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$. All p values in the Kaplan–Meier curves were corrected for multiple comparisons using the Benjamini & Hochberg method. The data fulfilled the preconditions/assumption of the above tests. Continuous CCS was normally distributed and variation was <1% between groups. All statistical analyses were performed using R-statistical software version 3.5.3 [27].

Data availability

The data used in this study are publicly available on the NIH website (https://gdc.cancer.gov/about-data/publications/pancanatlas).

Code availablity

R-code to reproduce the main and Supplementary results of this study are publicly available at https://github.com/arianlundberg/PANCAN.analysis.

Funding This work was supported by the Iris, Stig och Gerry Castenbäcks Stiftelse for cancer research (to NPT), the King Gustaf V Jubilee Foundation (NPT and JB), BRECT, the Swedish Cancer Society, the Cancer Society in Stockholm Personalised Cancer Medicine (PCM), the King Gustaf V Jubilee Foundation, the Swedish Breast Cancer Association (BRO) and the Swedish Research Council (J. Bergh). CMP was supported by funds from the NCI Breast SPORE program (P50-CA58223-09A1), by R01-CA195754-01, by the Susan G. Komen (SAC-160074) and the Breast Cancer Research Foundation.

Compliance with ethical standards

Conflict of interest JB has no conflict of interest related to the present work. Unrelated to the present work, he received research funding from Merck paid to Karolinska Institutet and from Amgen, Bayer, Pfizer, Roche and Sanofi-Aventis paid to Karolinska University Hospital. No personal payments. Payment from UpToDate for a chapter in breast cancer prediction paid to Asklepios Medicine AB. CMP is an equity stockholder, and consultant, for of BioClassifier LLC. CMP and JP are also listed as inventors on patents on the Breast PAM50 Subtyping assay. All remaining authors have declared no conflicts of interest.

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

References

1. Hartwell LH, Culotti J, Pringle JR, Reid BJ. Genetic control of the cell division cycle in yeast. Science. 1974;183:46–51.
2. Nurse P, Thuriaux P, Nasmyth K. Genetic control of the cell division cycle in the fission yeast Schizosaccharomyces pombe. Mol Gen Genet MGG. 1976;146:167–78.
3. Lee MG, Nurse P. Complementation used to clone a human homologue of the fission yeast cell cycle control gene cdc2. Nature. 1987;327:31–5.
4. Evans T, Rosenthal ET, Youngblom J, Distel D, Hunt T. Cyclin: a protein specified by maternal mRNA in sea urchin eggs that is destroyed at each cleavage division. Cell. 1983;33:389–96.
5. Hartwell LH, Kastan MB. Cell cycle control and cancer. Science. 1994;266:1821–8.
6. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell. 2011;144:646–74.
7. Hoadley KA, Yau C, Wolf DM, Cherniack AD, Tamborero D, Ng S, et al. Multiplatform analysis of 12 cancer types reveals molecular classification within and across tissues of origin. Cell. 2014;158:929–44.
8. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A. et al. Comprehensive characterization of cancer driver genes and mutations. Cell. 2018;173:371–85.e18.
9. Taylor AM, Shih J, Ha G, Gao GF, Zhang X, Berger AC. et al. Genomic and functional approaches to understanding cancer aneuploidy. Cancer Cell. 2018;33:676–89.e3.
10. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC. et al. Oncogenic signaling pathways in the cancer genome Atlas. Cell. 2018;173:321–37.e10.
11. Lundberg A, Lindström LS, Harrell JC, Falato C, Carlson JW, Wright PK, et al. Gene expression signatures and immunohistochemical subtypes add prognostic value to each other in breast cancer cohorts. Clin Cancer Res. 2017;23:7512–20.
12. Tobin NP, Lundberg A, Lindström LS, Harrell JC, Foukakis T, Carlsson L, et al. PAM50 provides prognostic information when applied to the lymph node metastases of advanced breast cancer patients. Clin Cancer Res. 2017;23:7225–31.

13. Chang F, Lee JT, Navolanic PM, Steelman LS, Shelton JG, Blalock WL, et al. Involvement of PI3K/Akt pathway in cell cycle progression, apoptosis, and neoplastic transformation: a target for cancer chemotherapy. Leukemia. 2003;17:590–603.
14. Chen J. The cell-cycle arrest and apoptotic functions of p53 in tumor initiation and progression. Cold Spring Harb Perspect Med. 2016;6:a026104.
15. Sotiriou C, Pusztai L. Gene-expression signatures in breast cancer. N. Engl J Med. 2009;360:790–800.
16. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.
17. Santos A, Wernersson R, Jensen LJ. Cyclebase 3.0: a multi-organism database on cell-cycle regulation and phenotypes. Nucleic Acids Res. 2015;43:D1140–4.
18. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43:D1079–85.
19. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell. 2018;173:400–16.e11.
20. Turner NC, Liu Y, Zhu Z, Loi S, Colleoni M, Loibl S, et al. Cyclin E1 expression and palbociclib efficacy in previously treated hormone receptor-positive metastatic breast cancer. J Clin Oncol. 2019;37:1169–78.
21. Sparano JA, Gray RJ, Ravdin PM, Makower DF, Pritchard KI, Albain KS, et al. Clinical and genomic risk to guide the use of adjuvant therapy for breast cancer. N. Engl J Med. 2019;380:2395–405.
22. Desmedt C, Haibe-Kains B, Wirapati P, Buyse M, Larsimont D, Bontempi G, et al. Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes. Clin Cancer Res. 2008;14:5158–65.
23. Asp M, Bergenstråhle J, Lundeberg J. Spatially resolved transcriptomes-next generation tools for tissue exploration. BioEssays. 2020;e1900221. https://onlinelibrary.wiley.com/doi/10.1002/bies.201900221.
24. Tobin NP, Lindström LS, Carlson JW, Bjöhle J, Bergh J, Wennmalm K. Multi-level gene expression signatures, but not binary, outperform Ki67 for the long term prognostication of breast cancer patients. Mol Oncol. 2014;8:741–52.
25. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. Cell. 2018;173:291–304. e6.
26. Carter SL, Cibulskis K, Helman E, McKenna A, Shen H, Zack T, et al. Absolute quantification of somatic DNA alterations in human cancer. Nat Biotechnol. 2012;30:413–21.
27. R Development Core Team (2008). R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.