Original Article

Classical pathology and mutational load of breast cancer – integration of two worlds

Jan Budczies,[1,2][†][*] Michael Bockmayr,[1][†] Carsten Denkert,[1,2] Frederick Klauschen,[1] Jochen K Lennerz,[3] Balázs Győrffy,[4,5,6] Manfred Dietel,[1,2] Sibylle Loibl,[7] Wilko Weichert[8,9,10] and Albrecht Stenzinger[8,11][*]

[1] Institute of Pathology, Charité University Hospital, Berlin, Germany
[2] German Cancer Consortium (DKTK), Berlin, Germany
[3] Massachusetts General Hospital/Harvard Medical School, Department of Pathology, Center for Integrated Diagnostics (CID), Boston, Massachusetts
[4] MTA TTK Lendület Cancer Biomarker Research Group, Budapest, Hungary
[5] 2nd Department of Pediatrics, Semmelweis University, Budapest, Hungary
[6] MTA-SE Pediatrics and Nephrology Research Group, Budapest, Hungary
[7] German Breast Group, Neu-Isenburg, Germany
[8] Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany
[9] National Center for Tumor Diseases (NCT), Heidelberg, Germany
[10] German Cancer Consortium (DKTK), Heidelberg, Germany
[11] Fellow of the NCT Heidelberg School of Oncology (NCT-HSO), Heidelberg, Germany

*Correspondence to: Albrecht Stenzinger, Institute of Pathology, University Hospital Heidelberg, Im Neuenheimer Feld 224, 69120 Heidelberg, Germany. e-mail: albrecht.stenzinger@med.uni-heidelberg.de or Jan Budczies, Institute of Pathology, Charité University Hospital, Charitéplatz 1, 10117 Berlin, Germany. e-mail: jan.budczies@charite.de

Abstract

Breast cancer is a complex molecular disease comprising several biological subtypes. However, daily routine diagnosis is still based on a small set of well-characterized clinico-pathological variables. Here, we try to link the two worlds of surgical pathology and multilayered molecular profiling by analyzing the relationships between clinico-pathological phenotypes and mutational loads of breast cancer. We evaluated the number of mutated genes with somatic non-silent mutations in different subgroups of breast cancer based on clinico-pathological, including immunohistochemical and tumour characteristics. The analysis was performed for a cohort of 687 primary breast cancer patients with mutational profiling, gene expression and clinico-pathological data available from The Cancer Genome Atlas (TCGA) project. The number of mutated genes was strongly positively associated with higher tumour grade ($p = 1.4e-14$) and with the different immunohistochemical and PAM50 molecular subtypes of breast cancer ($p = 1.4e-10$ and $p = 4.3e-10$, respectively). We observed significant associations ($|R| > 0.4$) between the abundance of mutated genes and expression levels of genes related to proliferation in the overall cohort and hormone receptor positive cohort, including the Recurrence Score gene signature (e.g., *MYBL2* and *BIRC5*). Specific mutated genes (*TP53*, *NCOR1*, *NF1*, *PTPRD* and *RB1*) were highly significantly associated with high loads of mutated genes. Multivariate analysis for overall survival (OS) revealed a worse survival for patients with high numbers of mutated genes (hazard ratio = 4.6, 95% CI: 1.0 – 20.0, $p = 0.044$). Here, we report a strong association of the number of mutated genes with immunohistochemical and PAM50 subtypes and tumour grade in breast cancer. We provide evidence that specific levels of the mutational load underlie different morphological and biological phenotypes, which collectively constitute the current basis of pathological diagnosis. Our study is a step towards genomics-informed breast pathology and will provide a basis for future studies in this field bridging the gap between morphology, tumour biology and medical oncology.

Introduction

Breast cancer is a complex molecular disease [1,2] and the most common cancer in women worldwide with each 8th to 9th woman affected during lifetime in the United States and Western Europe. While mortality is decreasing due to advances in both diagnostic and therapeutic approaches, breast cancer is still the second most common cause of death from cancer in women [3].

As demonstrated by a multitude of high-throughput molecular profiling studies [4–7] over the last 15 years spearheaded by the pioneering work of Perou *et al* [4], breast cancer is a heterogeneous disease comprising several biologically defined subtypes that are associated with distinct clinical behaviour and each require specific therapeutic strategies [8]. Currently, four subtypes (luminal A and luminal B, HER2-enriched and basal like/triple negative) are firmly established but more recent data even suggest a finer molecular classification in ten subtypes [9] whose clinical implications, however, still need to be explored. Complementing these advances, several multigene predictors have been developed to aid prediction of response towards adjuvant therapy and risk stratification of recurrence [10–12].

The current histomorphological typing system [13] still plays a significant role in the detection of rare subtypes but yields only limited clinically exploitable information for the most prevalent histological breast cancer type, invasive carcinoma of no special type (NST). With a shift to smaller tumours due to screening programs, tumour staging *per se* has only limited prognostic power [13,14] with recent studies suggesting that a refinement of the current staging system through inclusion of molecular profiles might prove beneficial [15,16]. The Elston–Ellis modification [17,18] of the Scarff–Bloom–Richardson grading system [19] is widely used to estimate outcomes but the semi-quantitative evaluation of morphological features rather than quantitative assessment of genetic parameters appears to introduce a considerable bias leaving grading results alone with a comparably limited impact on clinical decision making [13,14,20].

These developments and subsequent implementation of molecular classifications into routine diagnostics led to guidelines [21] which attribute a less important prognostic and particularly predictive power to classic morphology-based parameters, that is, typing, grading and staging [13,14,20–22].

This being said, it should be noted that, in principle, conventional morphology-based parameters, when applied in combination and evaluated by trained experts contribute to a reliable, valid and robust esti-

mation of the course of disease in many cancers including breast cancer [20]. Very surprisingly, however, studies investigating the actual genetic underpinnings of phenotype-based prognosticators and comparing these results with established molecular subtypes are still rare. A landmark study by Sotiriou *et al* [23] demonstrated histological grading to be associated with different gene expression signatures which also allowed stratifying the large set of grade 2 tumours into subgroups with different clinical outcome.

Interestingly, recent genomic profiling studies [24–31] suggested a substantial degree of intra- and inter-tumour heterogeneity fuelling selection processes during evolution of breast cancer [32,33], which may complicate prognostication as well as prediction and impede cancer precision medicine approaches [34]. In this context, it is worth recalling that just as the molecular phenotype, the morphological phenotype of the tumour, including tumour grade and tumour size is essentially a result of accumulated genetic aberrations over time, thereby reflecting tumour evolution at the phenotypic level.

Availability of a large dataset from The Cancer Genome Atlas (TCGA) enabled us to narrow in on these associations. Specifically, here, we investigated relations between the abundance of mutated genes and the morphological as well as biological phenotype of breast cancers.

Material and methods

Dataset

The study cohort (for details, see Table 1) included 687 breast cancer patients with both tumour exome sequencing and tumour grade data available. The primary variable under investigation was the number of mutated genes defined as genes carrying at least one somatic non-synonymous mutation in the coding region.

To obtain this quantity, we counted the number of genes with simple somatic mutations (substitutions, small insertions or deletions) in the coding region that led to a change in the amino acid sequence. Mutation data, clinico-pathological data and RNAseq data were downloaded from TCGA at https://tcga-data.nci.nih.gov/tcga/dataAccessMatrix.htm. Overall survival (OS) data with a limited mean follow-up time of 3.0 years and a limited median follow-up time of 2.0 years were avaiable. Seventy-four (10.8%) out of the 687 patients were followed until an event. Additionally, data on tumour grading and on tumours size (in cm) that were not available from TCGA were manually extracted from the corresponding pathology reports.

Characteristics	Group	Number of patients
Total	All patients	687 (100%)
Age	Median (q1-q3)	59 (49–68)
Size	Median (q1-q3)	2.5 (1.9–3.8)
N	N0	324 (47.2%)
	N+	351 (51.1%)
	NA	12 (1.8%)
G	G1	74 (10.8%)
	G2	319 (46.4%)
	G3	294 (42.8%)
Histopathology	Lobular	107 (15.6%)
	NST/other	579 (84.3%)
	NA	1 (0.2%)
Hormone receptor	HR−	148 (21.5%)
	HR+	532 (77.4%)
	NA	7 (1.0%)
HER2	HER2−	517 (75.3%)
	HER2+	113 (16.5%)
	NA	57 (8.3%)
Subtype	HR+/HER2−	405 (59.0%)
	HR+/HER2+	90 (13.1%)
	HR−/HER2−	111 (16.2%)
	HR−/HER2+	23 (3.4%)
	NA	58 (8.4%)
TP53	Mut	207 (30.1%)
	wt	480 (69.9%)
PIK3CA	Mut	212 (30.9%)
	wt	475 (69.1%)

q1 = first quartile, q3 = third quartile, NA = missing data

In the current study, we only used data sets that were published previously and labelled 'No restrictions; all data available without limitations' according to the TCGA publication guidelines (see http://cancergenome.nih.gov/publications/publicationguidelines, for details). We also extracted the PAM50 data supplied as supplementary data in [2].

With respect to the somatic mutation data, nonsynonymous mutations comprising insertions, deletions and substitutions in the coding region of genes were included. Structural rearrangements, aneuploidy, translocations or duplications as well as silent mutations were excluded from the analysis. For details visit https://wiki.nci.nih.gov/display/TCGA/Mutation+Annotation+Format.

Statistics

All statistical analyses and visualizations were performed using the statistical environment R including the R-packages clinfun, beeswarm, survival and binom [35–39].

The number of mutated genes was correlated with patient age and tumour size using Spearman correlations, with tumour grade using the Jonckheere–Terpstra test, with histopathological subtype, hormone receptor

(HR) status, HER2 status and molecular subtype using the exact Fisher test. p-values < 0.05 were considered statistically significant. Robust linear regression was executed using Huber's M-estimator as it is implemented in the function rlm of the R package MASS [40]. Correlations of the number of mutated genes with the mutation status of 23 recurrently mutated breast cancer genes [2] were investigated using Wilcoxon's rank test. p-values < 0.05 after Bonferroni-correction for the 23 genes were considered statistically significant.

Genome-wide gene expression data (generated by RNAseq) of breast cancer, HR+ breast cancer and HR− breast cancer were correlated with the number of mutated genes using the Spearman correlation coefficient. p-values < 0.05 after Bonferroni correlation for the total number of 20 531 genes were considered statistical significant. For visualization, fold changes (FC) between highly mutated tumours (22 or more mutations) and lowly mutated tumours (21 or less mutations) were calculated. Gene ontology (GO) categories were obtained from GO2MSIG [41].

Univariate and multivariate survival analysis was executed using the function coxph from the R package survival. p-values < 0.05 were considered significant. Prognostic evaluation of cut-off points for the number of mutated genes was executed as it is implemented in the Cutoff Finder software [42]. The mutation status of each of the 23 recurrently mutated breast cancer genes was correlated with patient age and tumour size using the Wilcoxon test, with tumour grade using the chi-squared test for trends, with histopathological subtype, HR status, HER2 status and molecular subtype using the exact Fisher test. p-values <0.05 after Bonferroni-correction for the 23 genes were considered statistically significant.

Results

We performed an integrated analysis of somatic mutations data, clinico-pathological data and RNAseq data that were publicly available from the TCGA project. Characteristics of the cohort comprising 687 patients are provided in Table 1. The primary variable under investigation was the number of genes carrying at least one non-silent somatic mutation (hereinafter simply termed 'mutated genes').

Associations of clinico–pathological parameters with mutational load. First, we correlated the number of mutated genes with the clinico-pathological characteristics of the tumour (Table 2, Figure 1). The number of mutated genes was strongly associated with tumour grade increasing from a median of 23 mutated genes

Table 2. Association of the number of mutated genes (non-silent somatic mutation) with clinico-pathological characteristics of breast cancer

Variable	Number of mutated genes (median)	*p*-value
Age	$R = 0.12$, slope $= 2.1$, intercept $= 26.0$	0.0016
1. Tumour size	$R = 0.17$, slope $= 1.6$ intercept $= 33.1$	0.00018
2. Nodal status	N0 = 34, N+ = 30	0.0048
3. Histopathology	ILC = 27, NST = 33	0.23
4. Tumour grade	G1 = 23, G2 = 27, G3 = 43	1.0e−14
5. Hormone receptor (HR)	HR+ = 28, HR− = 46	1.0e−10
6. HER2	HER2− = 31, HER2+ = 40	0.0066
7. Subtype (based on IHC)	HR+/HER2− = 27, HR+/Her2+ = 40, HR−/HER2− = 49, HR−/HER2+ = 41	1.0e−10
8. Subtype (based on PAM50)	Luminal A = 27, Luminal B = 39, Basal-like = 46, HER2-enriched = 50, Normal-like = 70	4.3e−10

The number of mutated genes correlated significantly with patient age, tumour size, nodal status, tumour grade, as well as immunohistochemical and PAM50 molecular subtypes. Associations with tumour grade, hormone receptor (HR) status and molecular subtype were much stronger compared with the associations with the other clinico-pathological parameters. Associations with patient age and tumour size were quantified using Spearman correlations and robust linear modelling. For patient age, the slope describes the increase in mutated genes per 10 years. For the tumour size, the slope describes the increase in mutated genes per 1cm increase in the maximum tumour dimension

in G1 tumors via 27 in G2 tumors to 43 in G3 tumours ($p = 1.4e−14$). When analysing the three components of tumour grade, nuclear pleomorphism (atypia) showed the strongest association with the number of mutated genes ($p = 5.5e−11$), but mitotic count and the degree of tubule formation also contributed significantly ($p = 4.8e−08$ and $p = 1.4e−03$, respectively). In addition, the number of mutated genes was strongly associated with the molecular subtype of breast cancer as determined by routine immunohistochemistry and the PAM50 classifier ($p = 1.4e−10$ and $p = 4.3e−10$, respectively). As shown in Table 2 and Figure 1, HR−/HER− breast cancer and HR−/HER2+ breast cancer harboured much more mutated genes compared with HR+/HER2− breast cancer. In contrast, we observed no statistically significant differences between the mutational loads of the two main histological subtypes, invasive lobular carcinoma (ILC) and NST (27 versus 23 mutated genes (median), $p = 0.23$). To narrow in on relations between histological tumour types and mutational load, we analyzed the abundance of mutated genes in all ILC (of any grade) and compared the results with the mutational load of NST carcinomas stratified according to grade (supplementary material Figure S1). While ILC harboured a mean of 27 mutated genes, G1–G3 NST displayed 23, 27 and 43 mutated genes, respectively ($p = 0.027$, $p = 0.66$, $p = 0.0011$). These data showed that ILC of any grade are fairly comparable with G2 NST with respect to their mutational load and ILC, G1 and G2 NST are in the lower range of mutational loads compared with G3 NST ($p = 9.12e−12$). Interestingly, ILC of any grade shared a high burden of mutated PIK3CA with G1–G3 NST. Highly mutated genes in ILC were *PIK3CA* (43.9%), CDH1 (56.1%), *MLL3* (7.5%), *FOXA1* and *TP53* (6.5% each), as well as *MAP3K1* and *GATA3* (4.7% each). 54.6% of NST

G1 showed mutated *PIK3CA*, followed by *MAP3K1* (16.4%), *GATA3* (14.6%), *MAP2K4* (10.4%), and *FOXP1* and *TP53* (3.6% each). With higher tumour grades, NST tumours with mutated *MAP2K4* (G2: 5.0%, G3:1.8%), *MAP3K1* (G2: 10.5%, G3: 3.2%) and *GATA3* (G2: 14.2%, G3: 5.6%) decreased while cases with *TP53* mutations strongly increased (G2: 19.7%, G3: 52.6%).

Further, we detected significant associations of the number of mutated genes with patient age ($p = 0.0016$), tumour size ($p = 0.00018$), and nodal stage ($p = 0.0048$). The median number of mutated genes in node positive (N+) tumours was 30 compared with 34 in N0 tumours. To quantify the strength of association with the continuous parameters, we executed robust linear regression of the number of mutated genes against patient age and against tumour size. Only 2.1 mutated genes were added per 10 years of the patient's life (slope, $p = 0.0012$), while 26.0 mutated genes (intercept, $p = 4.2e−09$) occurred independent of the age at diagnosis. Further, only 1.6 mutated genes were added when the tumour's maximum dimension extended by 1 cm (slope, $p = 0.0015$), while 33.1 mutated genes (intercept, $p = 5.8e−44$) occurred independent of the tumour size.

Finally, we executed the aforementioned analyses separately in HR+ breast cancer ($n = 532$) and in HR− breast cancer ($n = 148$). The number of mutated genes significantly correlated with tumour grade and HER2 status in HR+ breast cancer ($p = 4.8e−09$ and $p = 0.00039$), but not in HR− breast cancer. Interestingly, the correlation with patient age was stronger in HR+ breast cancer ($R = 0.2$) compared with HR− breast cancer ($R = 0.14$). Using estimates from robust linear regression, this corresponded to 3.4 mutated genes per 10 years life time occurring in HR+ tumours compared with 2.5 mutated genes per 10 years life time occurring in HR− tumours.
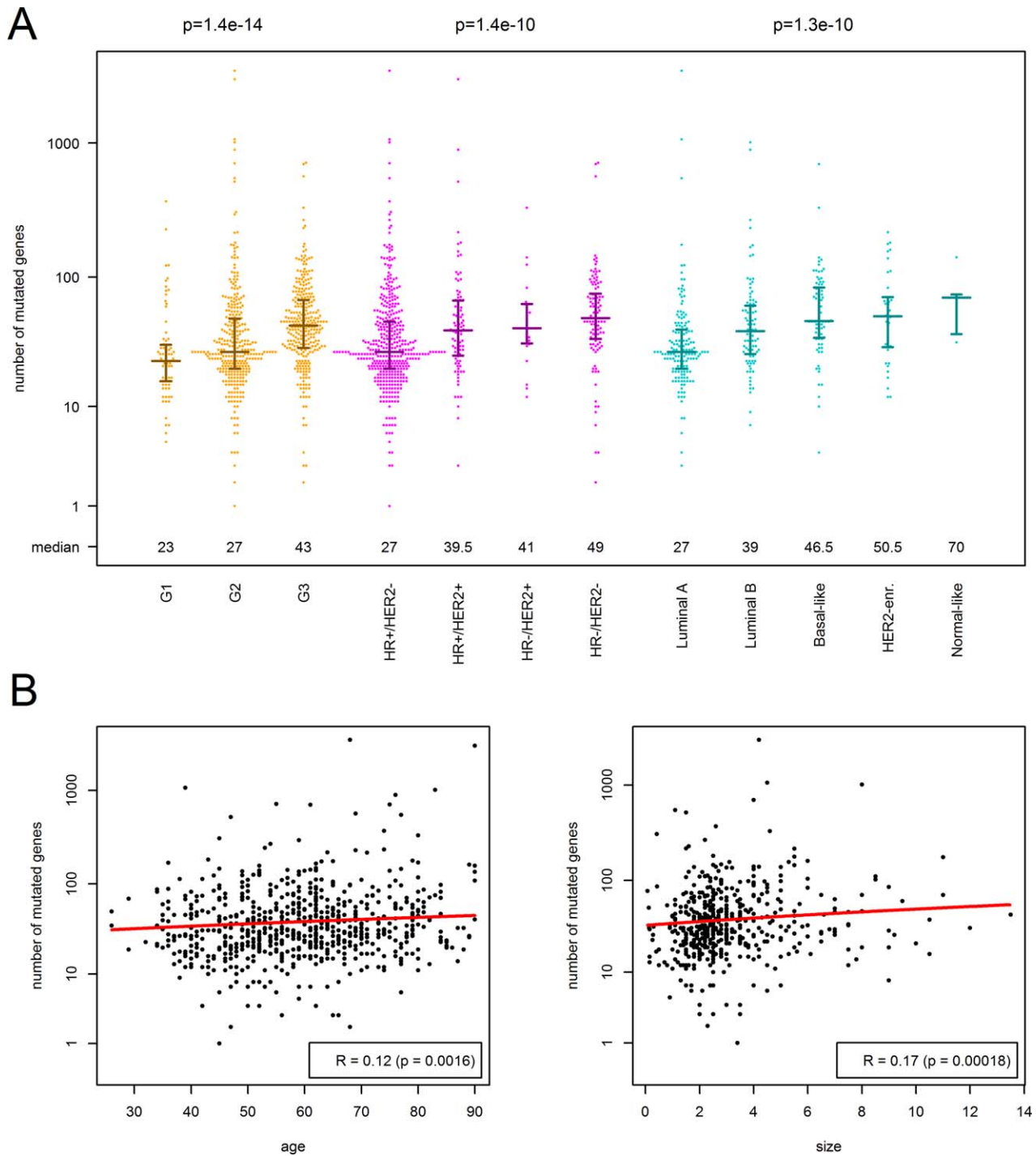
Figure 1. Association of the number of mutated genes (non-silent somatic mutation) with clinico-pathological characteristics of breast cancer. (A) In the beeswarm plot, each coloured dot represents a tumour. The bands indicate the first quartile, the median and the third quartile (*n* = median of mutated genes). The number of mutated genes was strongly associated with tumour grade, molecular subtyping by immunohistochemistry and SISH for ER, PR and HER2 as well as molecular subtyping by PAM50. (B) Scatterplots showing weak, but significant association of the number of mutations with patient age and tumour size in cm. The red curves represent the results of robust linear fitting. Not that the results of linear modeling appear as curves (not straight lines), because of the logarithmic scale of the *y*-axis.
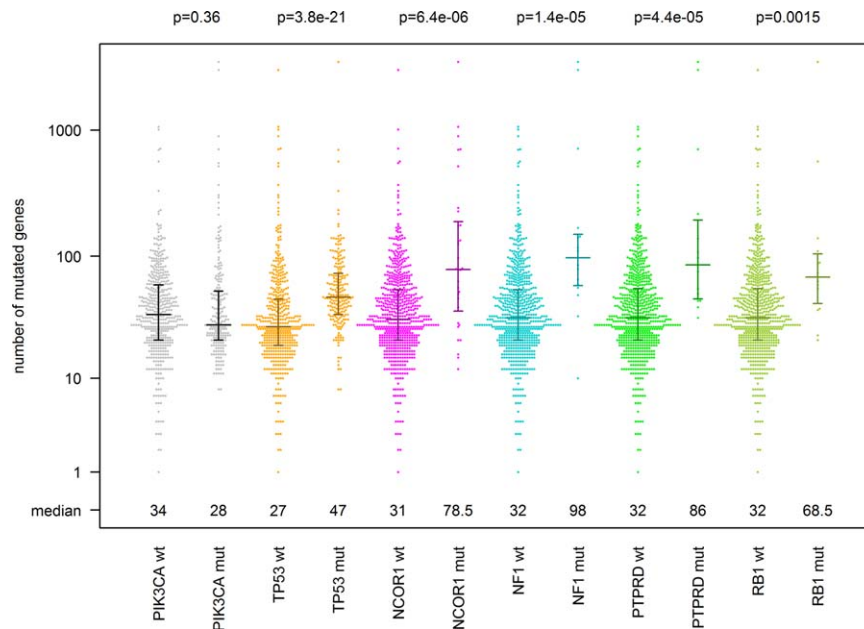
Figure 2. Association of the number of mutated genes (non-silent somatic mutation) with mutation status of genes frequently mutated in breast cancer. In the beeswarm plot, each coloured dot represents a tumour. The bands indicate the median including the first and third quartile (*n* = median of mutated genes). The number of mutated genes was not associated with *PIK3CA* mutation status, but strongly and highly significantly associated with mutated *TP53*. Moreover, the number of mutated genes was strongly and significantly associated with mutated *NCOR1*, *NF1*, *PTPRD* and *RB1*.

Associations of recurrently mutated genes with mutational load. Second, we correlated the number of mutated genes with the mutation status of each of the 23 genes that were reported by TCGA [2] as recurrently mutated in breast cancer (Figure 2, supplementary material Table S1). It turned out that the number of mutated genes was strongly associated with the mutation status of the tumour suppressor gene *TP53* (median of the number of mutated genes 47 in *TP53*-mutated tumours versus 27 in *TP53* wild-type tumours; $p = 3.8e-21$), while it was not associated with the mutation status of the oncogene *PIK3CA* ($p = 0.36$). Additionally, a significantly enhanced number of mutated genes was observed in, *NCOR1*-mutated, *NF1*-mutated, *PTPRD*-mutated and *RB1*-mutated tumours (Figure 2). In addition, we analyzed the correlation of the total mutational load with the mutation status of *BRCA1* and *BRCA2*. Note that, in this analysis, only somatic mutations were taken into account, as germline variant calling is currently not available for the TCGA BRCA data set. *BRCA1* somatic mutated tumours tended ($p = 0.058$) to harbour an enhanced total number of mutated genes ($n = 67$) compared with tumours without somatic *BRCA1* mutations ($n = 32$). *BRCA2* somatic mutated tumours had a significantly ($p = 0.0037$) enhanced total number of mutated genes

($n = 112$) compared with tumours without somatic *BRCA2* mutations ($n = 32$).

Associations of gene expression profiles with mutational load. Third, we analyzed the correlation of the gene expression landscape with the number of mutated genes (Figure 3, supplementary material Table S3). To this end, Spearman correlations were calculated between RNAseq data and the number of mutated genes. The number of genes whose expression showed a strong positive or negative correlation with the number of mutated genes ($|R| > 0.4$) was 44 in breast cancer (all $p < 8.8e-28$; Figure 4A), 11 in HR+ breast cancer (all $p < 4.7e-22$; Figure 4B) and 23 in HR− breast cancer (all; $p < 5.1e-07$; Figure 4D). In a functional analysis based on GO categories, we found the cellular process 'mitotic cell cycle (GO:0000278)' to be strongly and significantly enriched in the gene lists for breast cancer (20 genes, $p = 3.3e-19$) and for HR+ breast cancer (5 genes, $p = 1.0e-05$). When relaxing the correlation threshold to $|R| > 0.3$, larger numbers of 120 and 84 cell cycle genes belonged to the gene lists (significance of enrichment: $p = 2.6e-56$ and $p = 2.8e-52$, respectively). Moreover, we observed similar results for associations between the abundance of mutated genes and the five proliferation-related genes from
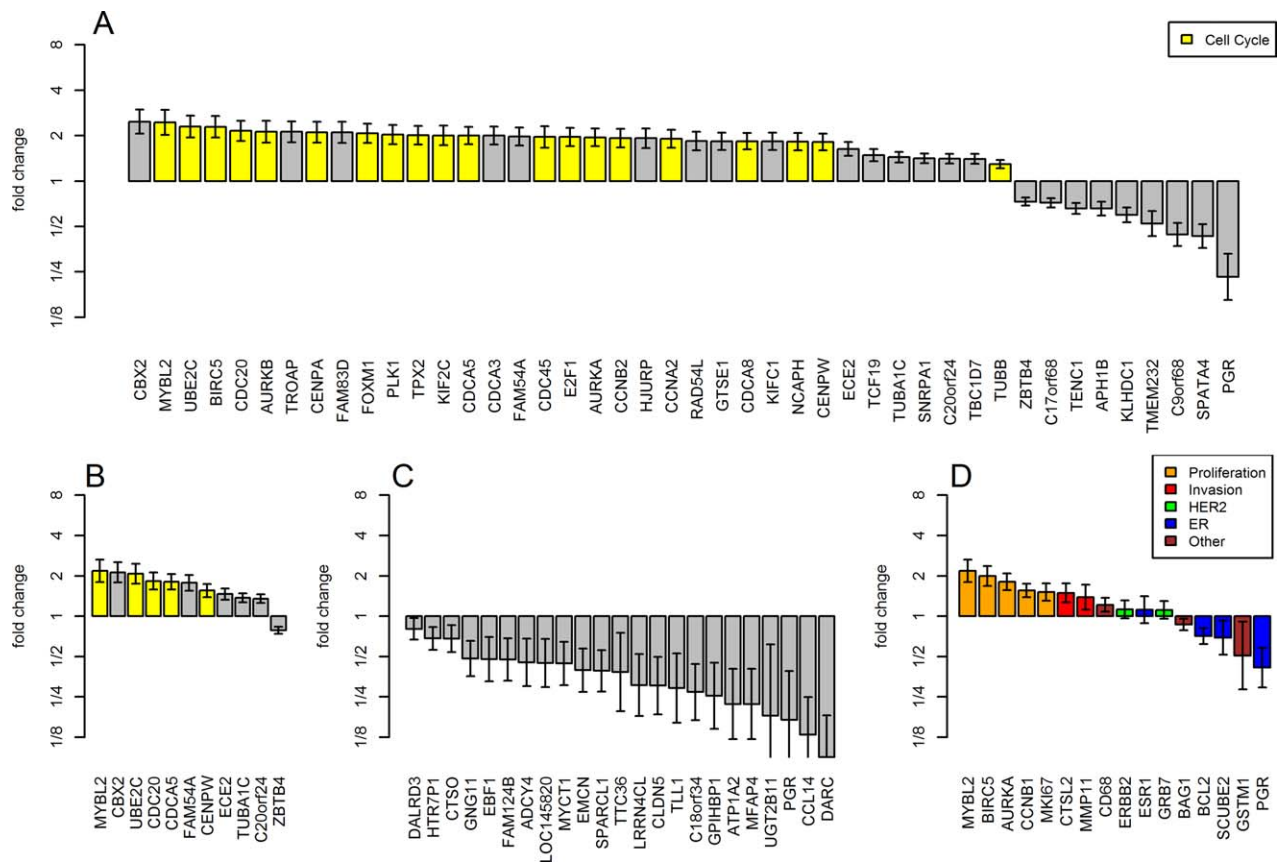
Figure 3. Gene expression FCs with 95% CIs between tumours with a high number of mutated genes (22 or more) and a low number of mutated genes (21 or less). Cell cycle genes [GO category 0000278: 'mitotic cell cycle'; (40)] are colored in yellow. (A) Analysis of breast cancer: FCs of 44 genes whose expression strongly correlated with the number of mutated genes ($|$Spearman-$R|>0.4$). (B) Analysis of HR+ breast cancer: FCs of 11 genes whose expression strongly correlated with the number of mutated genes ($|$Spearman-$R|>0.4$). (C) Analysis of HR− breast cancer: FCs of 23 genes whose expression strongly correlated with the number of mutated genes ($|$Spearman-$R|>0.4$). (D) Analysis of HR+ breast cancer: FCs of the 16 cancer genes included in the Recurrence Score assay.

the Recurrence Score gene signature [11] in HR+ breast cancers (all correlation $p < 2.4e-22$; Figure 4C, supplementary material Table S2). Taken together, in the overall cohort as well as in HR+ -driven tumours, we found especially upregulated cell cycle and proliferation-related genes to be associated with higher numbers of mutated genes. Specifically, *CDC20*, *MYBL2* and *BIRC5* (Survivin) as well as *MKI67*, which is routinely used for diagnostic work-up of breast cancers ranged among the top genes. In the gene expression data of HR− breast cancer, functional analysis did not detect any enriched GO category and all correlating genes were downregulated.

Associations of recurrently mutated genes with clinico-pathological parameters. Next, we correlated the mutation status of the 23 genes that were described as recurrently mutated in breast cancer [2] with tumour grade, molecular subtype and other clinico-pathological char-

acteristics of breast cancer (Figure 4). Mutation status of *PIK3CA*, *TP53*, *CDH1* and *CDKN1B* correlated significantly with tumour grade. Mutation status of *PIK3CA*, *TP53*, *GATA3* and *CDH1* correlated significantly with the molecular subtype. Further, we found a strong association between *CDH1* mutation status and the histopathological subtype, where 56.1% of the lobular cancers had *CHD1* mutations (G1: 57.9%, G2: 56.3%, G3: 50.0%) compared with only 2.1% of NST/other cancers (G1: 1.8%, G2: 2.1%, G3: 2.1%). Finally, we detected a moderate, but significant association of mutated *MLL3* ($p = 0.00046$) and FOXA1 ($p = 0.00097$) with patient age. Median age of patients harbouring these mutations was 63 and 67 compared with a median age of 59 in patients without these mutations.

Further, we investigated the correlation of somatic *BRCA1* and *BRCA2* mutation status with the clinico-pathological data. Mutations in *BRCA1* were
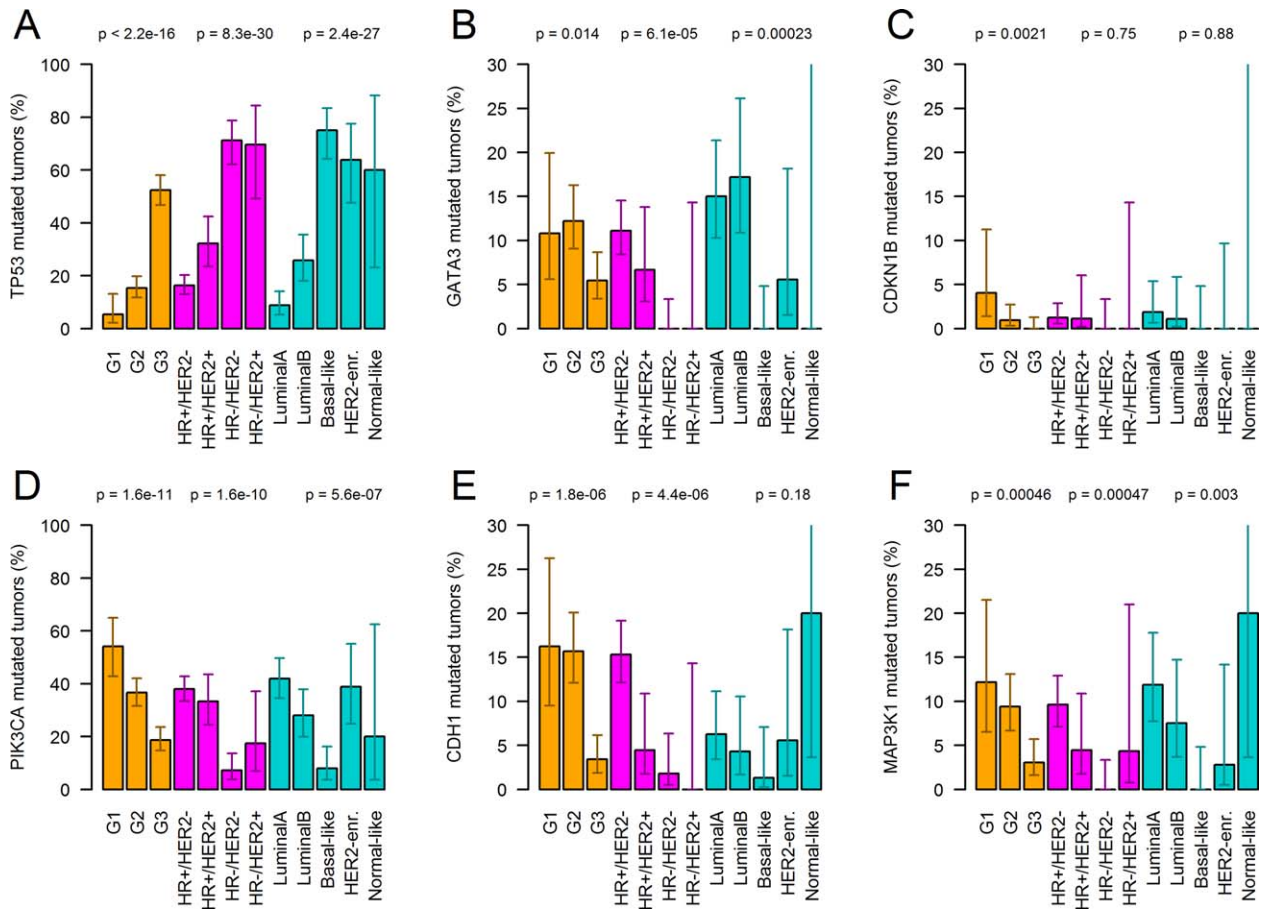
Figure 4. Association of the mutation status of specific genes with tumour grade and immunohistochemical as well as PAM50 sub-types (mutation rates including 95% confidence intervals). Genes with significant association after Bonferroni correction are shown (*TP53, GATA3, CDKN1B, PIK3CA, CDH1,* and *MAP3K1*). For example, the number of PIK3CA mutated tumours decreased from 54.3 via 36.7% to 18.7% with increasing tumour grade (G1, G2 and G3), while the number of *TP53* mutated tumours increased from 5.4 via 15.4% to 52.4%. Mutation status of *PIK3CA, TP53, GATA3, CDH1* and *MAP3K1* correlated significantly with the molecular subtype. *PIK3CA* mutations were much less frequent in triple-negative breast cancer (7.2%) compared with the other subtypes (all mutation rates >17.4%). *TP53* mutation rate was much higher in triple-negative (71.2%) and HR−/HER2+(69.6%) breast cancer compared with HR+/HER2−(16.3%) and HR+/HER2+(32.2%) breast cancer.

detected only in the basal-like molecular subtype (in 3.9% of these tumours), but absent in all other PAM50 subtypes ($p = 0.053$). BRCA2 mutations were found in the HER2-enriched subtype (8.3%) in the basal subtype (3.9%) and in the luminal A sub-type (1.2%), but absent in the other PAM50 subtypes ($p = 0.032$).

Associations of overall survival with mutational load. Finally, we investigated the correlation of outcome with the number of mutated genes (Figure 5). In univariate analysis of overall survival, the number of mutated genes was a negative prognostic marker with a hazard ratio of 1.18 per 100 mutated genes ($p = 0.00029$). In a multivariate analysis including correction for patient age, tumour size, nodal status,

histopathological type, tumour grade, HR status and HER2 status, the number of mutated genes was no longer significant. Next, we analyzed the prognostic power of a dichotomized version of the number of mutated genes. To this end, we made use of the Cut-off Finder [40] to study the impact of the choice of the cut-off point on differences in survival. In multivariate analysis, we detected a borderline-significant impact of the number of mutated genes on survival for cut-off points between 18.5 and 27.5 (Figure 4). An optimal cut-off point occurred when classifying tumours in those with 21 or less mutated genes and those with 22 or more mutated genes. This classification yielded a hazard ratio of 4.6 (95%CI: 1.0–20.0) and significance of $p = 0.044$ in multivariate analysis of overall survival.
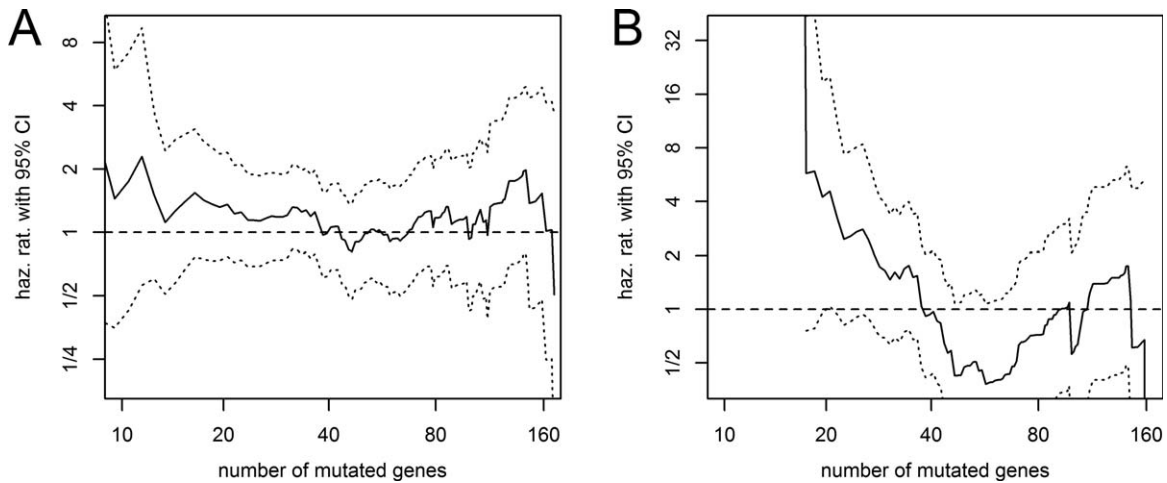
Figure 5. Association of overall survival with the total number of mutated genes (non-silent somatic mutation) in breast cancer. Hazard ratios (haz. rat.) are shown for a dichotomized version of the number of mutated genes that was varied along the x-axis. (A) In univariate analysis, no cut-off point between 10 and 180 yielded a significant association with survival. (B) In multivariate analysis including correction for patient age, tumour size, nodal status, histopathological type, tumour grade, hormone receptor status and HER2 status, a high number of mutations were borderline-significant associated with shorter survival for cut-off points between 18.5 and 27.5. The most significant association was obtained for a cut-off point of 21.5 with hazard ratio = 4.6 (1.0–10.0) and $p = 0.044$.

Discussion

Here, we show that the number of mutated genes, that is, the diversity of the gene pool putatively contributing to tumour evolution [34] is strongly associated with both morphological and biological features of breast cancer, which are commonly being used in routine diagnostics to guide clinical management and therapy worldwide. We also provide evidence for associations between the number of mutated genes and mutation status of specific genes as well as with expression levels of genes routinely used to assess the proliferative capacity of breast tumours (eg, Ki67, [43]) and to predict therapy response (eg, cancer gene signature of the Recurrence Score assay [11]).

Gene expression profiling studies combined with survival analysis [4–8] revealed that hormone receptor (HR)-positive (ie, luminal A and luminal B) and HR-negative tumours (ie, HER2-enriched and basal-like) are distinct diseases in both molecular and clinical terms. The intrinsic molecular subtypes are largely distinguished by gene sets measuring epithelial and basal differentiation as well as proliferation and the HER2 pathway. The robustness of the stratifying signatures has been demonstrated across several platforms [2,6,7]. In daily pathological diagnostics, however, immunohistochemistry is applied to facilitate approximation of these subtypes [43] and to stratify breast cancer patients into different risk groups and oncological therapies. A more recent combined analysis of genetic and RNA expression profiles of the largest breast cancer series to date comprising 2000 patients profiles by the META-BRIC study group [9], even revealed 10 different subgroups associated with different clinical outcomes; however, this has not entered clinical decision making yet.

In line with these findings, we observed significant differences in the relative abundance of mutated genes across molecular subtypes determined by immunohistochemistry and the PAM50 classifier: the median of mutated genes increased from 27 in HR+/HER2− tumours to 40 in HR+/HER2+ tumours, 41 in HR−/HER2+ and 49 in HR−/HER2− tumours and from 27 in Luminal A to 39 in Luminal B, 46.5 in Basal-like, 50.5 in HER2-enriched and 70 in normal-like tumours.

Recent large-scale molecular profiling studies [2,9] revealed considerable divergence of the main breast cancer subtypes on the genetic level. Interestingly, as shown by TCGA [2], luminal breast cancers are the most heterogeneous subtypes. Mutations frequently affected transcription factors of the ER pathway (GATA3, FOXA1 and RUNX1) and the PI3K cascade with mutations in PIK3CA as the most frequent genetic aberration followed by mutually exclusive mutations in PIK3R1, AKT and PTEN. Other frequent targets are the MAP-kinases MAP3K1 and MAP2K4 and epigenetic modulators, for example, genes encoding methyltransferases (eg, MLL3), demethyltransferases (eg, KDM6A) as well as several adenine-thymine-rich interactive domain-containing proteins (ARID1A and others). TP53 mutations occurred in

approx. 25.8% of luminal B and 8.8% of luminal A tumours. In contrast, basal-like breast cancer appears to harbour the most unique and distinctive genomic profile. They are primarily characterized by genomic instability that is putatively caused by the loss of function of three tumour suppressor genes and cell-cycle regulators by mutations or deletions: *TP53* (75%), *RB1* (4%), and *BRCA1/2* (7.9%). Similar to basal-like breast cancers, HER2-enriched subtypes showing a certain relation to luminal subtypes also harbour a high frequency of *TP53* (63.9%) that is accompanied by *PIK3CA* mutations (38.9%).

In accord with these observations, extended analysis of the whole cohort showed a significantly high prevalence of mutated *TP53* in TNBC and HER2-enriched breast cancer, while mutated *CDH1* and *GATA3* occur significantly more frequently in the HR+ subtype and the HR+/HER2− subtype, respectively. We also noted that while *PIK3CA* and *CDH1* mutations were significantly enriched in tumours with lower tumour grades, *TP53* and *CDKN1B* were significantly associated with higher tumour grades. However, when analyzing *CDH1* status within histological subtypes, we did not observe any significant correlations with tumour grade. As already reported by other investigators [2,44,45], mutated *CDH1* occured almost exclusively in lobular breast cancer.

Interestingly, we did not observed significant associations between the abundance of mutated genes and several of the frequently mutated breast-cancer related genes briefly summarized above. However, we identified five mutated genes (*TP53*, *NCOR1*, *PTRD*, *NF1* and *RB1*), which showed strong associations to high numbers of mutated genes independent of the tumour subtype. As expected, non-silent mutations in TP53 and RB1, which have been causally linked to genomic instability of tumours [46,47], are strongly correlated with the number of mutated genes in breast cancer. Moreover, our data show that high numbers of mutated genes are strongly associated with mutant *NCOR1*, which is mutated in ~4% of breast cancer cases [2]. Although we cannot provide evidence for a causal relationship, it is of note that *NCOR1* is a gene encoding a transcriptional co-regulator serving as a linker for transcription factors and chromatin-modifying enzymes and has been implicated in maintaining genomic stability via HDAC3 [48]. Pointing towards a similar direction, the mutant variant of *PTPRD* (genetic aberrations in ~7% of breast cancers [2]) encoding the protein tyrosine phosphatase receptor type delta was also strongly associated with high numbers of mutated genes in breast cancer. *PTPRD* has been shown to act as tumour suppressor [49] and impairs function of

Aurora A [50], a protein that has been shown to modulate genomic stability [51]. Last but not least, the mutated variant of the putative breast cancer oncogene *NF1* (genetic alterations in ~5% of breast cancers [2]) correlated with a high abundance of mutated genes in our study. Neurofibromin1 acts as a negative regulator of *RAS* by stimulating GTPase activity of *RAS* [52] and impaired *NF1* function has been linked to increased mutations frequencies in mice [53]. Taken together, our observational data indicate that it may be worthwhile to investigate the specific functional relevance of these five genes in breast cancer biology and maintenance of genome integrity in future studies.

In addition, we sought to identify differentially expressed genes that are associated with the number of mutated genes in breast cancer. Interestingly, we found many upregulated cell cycle-related proteins to be related to high numbers of mutated genes in the overall cohort. Among the top upregulated genes were *MYBL2*, *BIRC5* and *AURKA*, which are also part of the cancer genes interrogated by the Recurrence Score assay [11]. Narrowing in on this, we observed that particularly the proliferation gene signature of the Recurrence score shows a strong correlation with the mutation rate in HR+ breast cancer subtypes. A broader analysis of the HR+ cancers corroborated *MYBL2* and identified other upregulated cell-cycle associated genes, such as *PLK1* and *CDC20*. These data suggest that the link between proliferation and abundant mutations is largely restricted to HR+ tumours and does not pertain to HR− tumours. It is tempting to speculate that this proliferation signature indicates that increased tumour cell replication and associated stochastically higher mutation rate in luminal tumours fuels the pool of mutated genes thereby advancing tumour heterogeneity and evolutionary selection processes [34].

Remarkably, we found that tumour grade, that is, phenotypes indicating the degree of cellular differentiation, but not the two main breast cancer types, that is, NST and ILC, are strongly associated with different numbers of mutated genes. Additionally, we identified a significant but comparably weaker correlation of the mutational load with two other measures reflecting the temporal development of the tumour, that is, age at diagnosis and tumour size. Specifically, the expected number of acquired mutated genes was estimated as 2.1 per 10 years and 1.6 per 1 cm, respectively. Hypothesizing that the abundance of mutated genes is a measure of genetic heterogeneity of the tumour [34], we here provide evidence that particularly pathological grading of breast cancer is in fact a microscopic read-out of tumour

heterogeneity, which indicates different biological and clinical behaviour of the tumour. Several studies either sequencing tumour bulks or single tumour cells have demonstrated that breast cancer subtypes exhibit considerable spatial and temporal heterogeneity on the genetic level within the primary, during metastatic progression and in patient-derived xenografts [24–31]. This genomic diversity within breast cancers is a result of but also facilitates cellular evolution [34] allowing the tumour to dynamically adapt to external or internal stimuli as originally conceptualized by Nowell in 1976 [54].

While we and others [31,33,34,54] hypothesize that this diverse gene pool may confer survival advantage to the tumour, to our surprise and counter-intuitive, local spread to lymph nodes was associated with slightly but significantly lower numbers of mutated genes compared with node negative tumours. Currently, we are unable explain these results satisfactorily and future studies are needed to investigate these relations further. Although data on follow-up times of the cohort are rather short and limited, and hence, limit analysis as well as data interpretation, our survival data analysis indicates that increasing numbers of mutated genes *per se* are significantly associated with increased hazard ratios independently of other clinico-pathological parameters. Hence, our data may suggest that the abundance of mutated genes has a certain relation with metastatic progression ultimately leading to tumour-related death [55–57]. Further studies are warranted to substantiate our findings and conclusions.

To sum up, we here provide evidence for significant associations between the genotype and morphological as well biological phenotypes of breast cancer. Specifically, while histological subtypes do not display specific numbers of non-synonymous somatic mutations, immunohistochemically defined subtypes of breast cancer are associated with different amounts of mutated genes increasing from HR+/HER2− to HR+/HER2+, HR−/HER2+ and HR−/HER2− breast cancers. Moreover, we observed significant correlations with the expression level of genes involved in cellular proliferation as well as with specific breast-cancer related genes. Interestingly, the increasing abundance of mutated genes displayed strong correlations with increasing tumour grades and also with tumour size and age at diagnosis albeit to a lower degree. The three main implications of these observations are: 1) a critical reappraisal of tumour grading that warrants further investigation as the tumour grade assessed by conventional light-microscopy for prognostication of the course of disease may reflect genetic tumour hetero-

geneity and in turn possibly the evolutionary potential of tumour cells, 2) different molecular subtypes of breast cancer may harbour a different capacity for tumour evolution due to different abundances of mutated genes and 3) we provide evidence for a strong relation between the proliferative capacity as well as specific mutated genes and the number of genes harbouring somatic non-silent mutations in breast cancer.

In conclusion, we believe that our study is a step further towards genomics-informed breast pathology and will provide a basis for future studies in this field bridging the gap between morphology, tumour biology and medical oncology.

List of abbreviations

AFF2	AF4/FMR2 family member 2
ARID1A	AT-rich interactive domain-containing protein 1A
AURKA/STK15	Aurora A kinase
BIRC5	Survivin
CBFB	Core-binding factor subunit beta
CDH1	Cadherin-1
CDC20	cell-division cycle protein 20
CTCF	CCCTC-binding factor
ER	Estrogen receptor
FC	Fold change
GATA3	Transacting T-cell-specific transcription factor 3
HER2	Human epidermal growth factor receptor 2
HR	Hormone receptor
ILC	Invasive lobular carcinoma
MAP3K1	Mitogen-activated protein kinase kinase kinase 1
MAP2K4	mitogen-activated protein kinase kinase 4
MLL3	Histone-lysine N-methyltransferase 3
MYBL2	Myb-related protein B
NCOR1	Nuclear receptor corepressor 1
NF1	Neurofibromin 1
NST	Invasive carcinoma of no special type
TNBC	Triple negative breast cancer
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha
PLK1	Polo-like kinase
PTEN	Phosphatase and tensin homolog
PTPRD	Receptor-type tyrosine-protein phosphatase delta
PR	Progesteron receptor

RB1	retinoblastoma 1
RUNX1	Runt-related transcription factor 1
SF3B1	Splicing factor 3B subunit 1
SPC24	kinetochore protein SPc 24
TBX3	T-box transcription factor
TP53	tumour suppressor 53

Acknowledgements

Author contributions

AS and JB conceived and designed the study and supervised experiments. JB, MB, FK, JKL, SB, GB, MD, CD, WW and AS analyzed sequencing data and clinicopathological parameters. AS and JB wrote the manuscript with contributions from MB, FK, SB, CD, JKL, WW, and SL. All authors read and approved the final manuscript.

References

1. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, *et al*. The landscape of cancer genes and mutational processes in breast cancer. *Nature* 2012; **486**: 400–404.

2. Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61–70.

3. DeSantis C, Ma J, Bryan L, Jemal A. Breast cancer statistics, 2013. *CA Cancer J Clin* 2014; **64**: 52–62

4. Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, *et al*. Molecular portraits of human breast tumours. *Nature* 2000; **406**: 747–752.

5. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, *et al*. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci USA* 2001; **98**: 10869–10874.

6. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, *et al*. Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*. 2003; **100**: 8418–8423.

7. Wirapati P, Sotiriou C, Kunkel S, Farmer P, Pradervand S, Haibe-Kains B, *et al*. Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res* 2008; **10**: R65.

8. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *Lancet* 2011; **378**: 1812–1823.

9. Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 2012; **486**: 346–352.

10. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, *et al*. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med* 2002; **347**: 1999–2009.

11. Paik S, Shak S, Tang G, Kim C, Baker J, Cronin M, *et al*. A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *N Engl J Med* 2004; **351**: 2817–2826.

12. Kronenwett R, Bohmann K, Prinzler J, Sinn BV, Haufe F, Roth C, *et al*. Decentral gene expression analysis: analytical validation of the Endopredict genomic multianalyte breast cancer prognosis test. *BMC Cancer* 2012; **12**: 456.

13. Viale G. The current state of breast cancer classification. *Ann Oncol* 2012; **23**(Suppl 10): x207–10.

14. Singletary SE, Allred C, Ashley P, Bassett LW, Berry D, Bland KI, *et al*. Revision of the American Joint Committee on Cancer staging system for breast cancer. *J Clin Oncol* 2002; **20**: 3628–3636.

15. Mittendorf EA, Jeruss JS, Tucker SL, Kolli A, Newman LA, Gonzalez-Angulo AM, *et al*. Validation of a novel staging system for disease-specific survival in patients with breast cancer treated with neoadjuvant chemotherapy. *J Clin Oncol* 2011; **29**: 1956–1962.

16. Bagaria SP, Ray PS, Sim MS, Ye X, Shamonki JM, Cui X, *et al*. Personalizing breast cancer staging by the inclusion of ER, PR, and HER2. *JAMA Surg* 2014; **149**: 125–129.

17. Ellis IO, Galea M, Broughton N, Locker A, Blamey RW, Elston CW. Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. *Histopathology* 1992; **20**: 479–489.

18. Elston CW, Ellis IO. Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology* 1991; **19**: 403–410.

19. Bloom HJ, Richardson WW. Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *Br J Cancer* 1957; **11**: 359–377.

20. Rakha EA, Reis-Filho JS, Baehner F, Dabbs DJ, Decker T, Eusebi V, *et al*. Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Res* 2010; **12**: 207.

21. Goldhirsch, A, Wood WC, Coates AS, Gelber RD, Thürlimann B, Senn HJ, *et al*. Strategies for subtypes—dealing with the diversity of breast cancer: highlights of the St Gallen International Expert Consensus on the Primary Therapy of Early Breast Cancer 2011. *Ann Oncol* 2011; **22**: 1736–1747.

22. Rakha EA, Ellis IO. Modern classification of breast cancer: should we stick with morphology or convert to molecular profile characteristics. *Adv Anat Pathol* 2011; **18**: 255–267.

23. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, Smeds J, *et al*. Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst* 2006; **98**: 262–272.

24. Shah SP, Morin RD, Khattra J, Prentice L, Pugh T, Burleigh A, *et al*. Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* 2009; **461**: 809–813.

25. Shah SP, Roth A, Goya R, Oloumi A, Ha G, Zhao Y, *et al*. The clonal and mutational evolution spectrum of primary triple-negative breast cancers. *Nature* 2012; **486**: 395–399.

26. Wang Y, Waters J, Leung ML, Unruh A, Roh W, Shi X, et al. Clonal evolution in breast cancer revealed by single nucleus genome sequencing. *Nature* 2014; **512**: 155–160.

27. Eirew P, Steif A, Khattra J, Ha G, Yap D, Farahani H, et al. Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution. *Nature* 2014; **518**: 422–426.

28. Nik-Zainal S, Alexandrov LB, Wedge DC, Van Loo P, Greenman CD, Raine K, et al. Breast cancer working group of the international cancer genome consortium. Mutational processes molding the genomes of 21 breast cancers. *Cell* 2012; **149**: 979–993.

29. Nik-Zainal S, Van Loo P, Wedge DC, Alexandrov LB, Greenman CD, Lau KW, et al. Breast cancer working group of the international cancer genome consortium. The life history of 21 breast cancers. *Cell* 2012; **149**: 994–1007.

30. Almendro V, Cheng YK, Randles A, Itzkovitz S, Marusyk A, Ametller E, et al. Inference of tumor evolution during chemotherapy by computational modeling and *in situ* analysis of genetic and phenotypic cellular diversity. *Cell Rep* 2014; **6**: 514–527.

31. Almendro V, Kim HJ, Cheng YK, Gönen M, Itzkovitz S, Argani P, et al. Genetic and phenotypic diversity in breast tumor metastases. *Cancer Res* 2014; **74**: 1338–1348.

32. Campbell LL, Polyak K. Breast tumor heterogeneity: cancer stem cells or clonal evolution? *Cell Cycle* 2007; **6**: 2332–2338.

33. Martelotto LG, Ng CK, Piscuoglio S, Weigelt B, Reis-Filho JS, et al. Breast cancer intra-tumor heterogeneity. *Breast Cancer Res* 2014; **16**: R48.

34. Burrell RA, McGranahan N, Bartek J, Swanton C, et al. The causes and consequences of genetic heterogeneity in cancer evolution. *Nature*. 2013; **501**: 338–345.

35. R Development Core Team. 2014. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org/. last accessed on May 20, 2015.

36. Seshan VE. 2014. Clinfun: clinical trial design and data analysis functions. R package version 1.0.6. Available at: http://CRAN.R-project.org/package=clinfun. last accessed on May 20, 2015.

37. Eklund A. 2013. Beeswarm: the bee swarm plot, an alternative to stripchart. R package version 0.1.6. Available at: http://CRAN. R-project.org/package=beeswarm. last accessed on May 20, 2015.

38. Therneau TM, Grambsch PM. *Modeling Survival Data: Extending the Cox Model*. Springer: New York, 2000; ISBN 0-387-98784-3.

39. Dorai-Raj S. 2014. Binom: binomial confidence intervals for several parameterizations. R package version 1.1-1. Available at: http://CRAN.R-project.org/package=binom. last accessed on May 20, 2015.

40. Venables WN, Ripley BD. *Modern Applied Statistics with S* (4th edn). Springer: New York, 2002; ISBN 0–387-95457-0.

41. Powell JA. GO2MSIG, an automated GO based multi-species gene set generator for gene set enrichment analysis. *BMC Bioinformatics* 2014; **15**: 146.

42. Budczies J, Klauschen F, Sinn BV, Györffy B, Schmitt WD, Darb-Esfahani S, et al. Cutoff Finder: a comprehensive and straightforward web application enabling rapid biomarker cutoff optimization. *PLoS One*. 2012; **7**: e51862.

43. Cornejo KM, Kandil D, Khan A, Cosar EF, et al. Theranostic and molecular classification of breast cancer. *Arch Pathol Lab Med* 2014; **138**: 44–56

44. Bertucci F, Orsetti B, Nègre V, Finetti P, Rougé C, Ahomadegbe JC, et al. Lobular and ductal carcinomas of the breast have distinct genomic and expression profiles. *Oncogene* 2008; **27**: 5359–5372.

45. Petridis C, Shinomiya I, Kohut K, Gorman P, Caneppele M, Shah V, et al. Germline CDH1 mutations in bilateral lobular carcinoma in situ. *Br J Cancer* 2014; **110**: 1053–1057.

46. Negrini S, Gorgoulis VG, Halazonetis TD. Genomic instability—an evolving hallmark of cancer. *Nat Rev Mol Cell Biol* 2010; **11**: 220–228.

47. Hernando E, Nahle Z, Juan G, Diaz-Rodriguez E, Alaminos M, Hemann M, et al. Rb inactivation promotes genomic instability by uncoupling cell cycle progression from mitotic control. *Nature* 2004; **430**: 797–802.

48. Bhaskara S, Knutson SK, Jiang G, Chandrasekharan MB, Wilson AJ, Zheng S, et al. Hdac3 is essential for the maintenance of chromatin structure and genome stability. *Cancer Cell* 2010; **18**: 436–447.

49. Veeriah S, Brennan C, Meng S, Singh B, Fagin JA, Solit DB, et al. The tyrosine phosphatase PTPRD is a tumor suppressor that is frequently inactivated and mutated in glioblastoma and other human cancers. *Proc Natl Acad Sci USA* 2009; **106**: 9435–9440.

50. Meehan M, Parthasarathi L, Moran N, Jefferies CA, Foley N, Lazzari E, et al. Protein tyrosine phosphatase receptor delta acts as a neuroblastoma tumor suppressor by destabilizing the aurora kinase A oncogene. *Mol Cancer* 2012; **11**: 6.

51. Ertych N, Stolz A, Stenzinger A, Weichert W, Kaulfuβ S, Burfeind P, et al. Increased microtubule assembly rates influence chromosomal instability in colorectal cancer cells. *Nat Cell Biol* 2014; **16**: 779–791.

52. Wallace MD, Pfefferle AD, Shen L, McNairn AJ, Cerami EG, Fallon BL, et al. Comparative oncogenomics implicates the neurofibromin 1 gene (NF1) as a breast cancer driver. *Genetics* 2012; **192**: 385–396.

53. Garza R, Hudson RA 3rd, McMahan CA, Walter CA, Vogel KS, et al. A mild mutator phenotype arises in a mouse model for malignancies associated with neurofibromatosis type 1. *Mutat Res* 2007; **615**: 98–110.

54. Nowell PC. The clonal evolution of tumor cell populations. *Science* 1976; **194**: 23–28.

55. Baccelli I, Schneeweiss A, Riethdorf S, Stenzinger A, Schillert A, Vogel V, et al. Identification of a population of blood circulating tumor cells from breast cancer patients that initiates metastasis in a xenograft assay. *Nat Biotechnol* 2013; **31**: 539–544.

56. Baccelli I, Stenzinger A, Vogel V, Pfitzner BM, Klein C, Wallwiener M, et al. Co-expression of MET and CD47 is a novel prognosticator for survival of luminal breast cancer patients. *Oncotarget* 2014; **5**: 8147–8160.

57. Budczies J, von Winterfeld M, Klauschen F, Bockmayr M, Lennerz JK, Denkert C, et al. The landscape of metastatic progression patterns across major human cancers. *Oncotarget* 2015; **6**: 570–583.