Breast Cancer Research

RESEARCH

Open Access

Integrative multi-omic sequencing reveals the MMTV-Myc mouse model mimics human breast cancer heterogeneity

Carson D. Broeker[1], Mylena M. O. Ortiz[2], Michael S. Murillo[3,4] and Eran R. Andrechek[5*]

Abstract

Background Breast cancer is a complex and heterogeneous disease with distinct subtypes and molecular profiles corresponding to different clinical outcomes. Mouse models of breast cancer are widely used, but their relevance in capturing the heterogeneity of human disease is unclear. Previous studies have shown the heterogeneity at the gene expression level for the MMTV-Myc model, but have only speculated on the underlying genetics.

Methods Tumors from the microacinar, squamous, and EMT histological subtypes of the MMTV-Myc mouse model of breast cancer underwent whole genome sequencing. The genomic data obtained were then integrated with previously obtained matched sample gene expression data and extended to additional samples of each histological subtype, totaling 42 gene expression samples. High correlation was observed between genetic copy number events and resulting gene expression by both Spearman's rank correlation coefficient and the Kendall rank correlation coefficient. These same genetic events are conserved in humans and are indicative of poor overall survival by Kaplan–Meier analysis. A supervised machine learning algorithm trained on METABRIC gene expression data was used to predict the analogous human breast cancer intrinsic subtype from mouse gene expression data.

Results Herein, we examine three common histological subtypes of the MMTV-Myc model through whole genome sequencing and have integrated these results with gene expression data. Significantly, key genomic alterations driving cell signaling pathways were well conserved within histological subtypes. Genomic changes included frequent, co-occurring mutations in KIT and RARA in the microacinar histological subtype as well as SCRIB mutations in the EMT subtype. EMT tumors additionally displayed strong KRAS activation signatures downstream of genetic activating events primarily ascribed to KRAS activating mutations, but also FGFR2 amplification. Analogous genetic events in human breast cancer showed stark decreases in overall survival. In further analyzing transcriptional heterogeneity of the MMTV-Myc model, we report a supervised machine learning model that classifies MMTV-Myc histological subtypes and other mouse models as being representative of different human intrinsic breast cancer subtypes.

Conclusions We conclude the well-established MMTV-Myc mouse model presents further opportunities for investigation of human breast cancer heterogeneity.

Keywords Myc, Breast cancer, Mammary tumor, Whole genome sequencing, Transcriptomics, Machine learning, Heterogeneity, Multi-omics, Comparative genomics

*Correspondence:
Eran R. Andrechek
andrech1@msu.edu
Full list of author information is available at the end of the article

Background

MYC is a master transcriptional regulator that is amplified in approximately 15–20% of breast cancers and overexpressed in up to 35% of breast cancers [1]. All MYC family genes (c-MYC, N-MYC, L-MYC, B-MYC) contain basic helix-loop-helix (bHLH) domains, which allow for heterodimerization with MYC-associated factor X (MAX) to bind consensus E-box sequence motifs on core gene promoter regions [2]; this allows for the initiation of transcription of MYC responsive genes, including regulation of proliferation, cell growth, differentiation, nucleotide biosynthesis, DNA replication, RNA levels, and apoptosis [3, 4]. MYC amplification is often enriched in high-grade breast cancers [5, 6], triple-negative breast cancers [7], and basal-like breast cancers [8]. This is supported by evidence showing that MYC occupies the promoters of most active genes in tumorigenesis and that MYC acts as a non-specific amplifier of whole cell gene expression [9]. Triple-negative breast cancer (TNBC) and basal-like breast cancer subtypes carry poor prognostic clinical outcomes relative to other breast cancer subtypes [10]. In mouse models, several groups have demonstrated that overexpression of MYC alone is sufficient for the development of spontaneous mammary tumors with high penetrance [11–13].

There is a clear, concerted interest to develop mouse models of breast cancer that recapitulate the heterogeneous features of human breast cancer to better understand disease dynamics. In response, numerous mouse models of breast cancer have been created, characterized by a wide variety of genetic drivers including constitutive overexpression of endogenous oncogenes specifically in the mouse mammary gland (MMTV-Neu, MMTV-Cyclin D1, MMTV-Akt1 [14]), conditional overexpression of oncogenes in the mammary gland (MMTV-rtTA/TetO-NeuNT [15], MTB/TWNT [16]), nonfunctional or conditional loss of tumor-suppressor genes (Stat1$^{-/-}$, MMTV-Cre/BRCA1$^{fl/fl}$) [14], or overexpression of exogenous transforming oncoproteins (MMTV-PyMT [17]). Various mouse models of breast cancer have had their gene expression profiles characterized [18, 19], with most models clustering largely within one human intrinsic breast cancer subtype. While many mouse models have been generated that model a specific aspect of human breast cancer biology, few realize the full spectrum of heterogeneity present within human breast tumors or accurately model clinical features [20]. Previous studies revealed the MMTV-Myc model of breast cancer mimics many human disease parameters, including substantial histological and transcriptional heterogeneity 13]. These analyses revealed a close association between the transcriptional profile and histological subtype among MMTV-Myc tumors. Other studies have also shown that tumors derived from different mouse models that share the same histological patterns cluster more tightly transcriptionally than they do with tumors of the same genotype with different histological features [18]. Importantly, MMTV-Myc samples have one of the most varied gene expression patterns amongst models, with distinct subsets of MMTV-Myc samples clustering closely with all intrinsic human breast cancer subtypes [18].

While the contributions of the tumor microenvironment, epigenetics, tumor metabolome, immune composition, and other factors have been highlighted recently for their roles in cancer [21–23], most often cancers are driven by genetic aberrations by activation of oncogenes and inactivation of tumor-suppressor genes. The advent of anti-estrogenic compounds such as tamoxifen in the treatment of hormone-receptor positive breast cancers [24] and the monoclonal antibody trastuzumab to treat HER2+ breast cancer [25] have demonstrated the utility in targeting breast cancer with therapies based on genomic, transcriptomic, and immunohistochemical data. In realization of the complex molecular aberrations driving all cancers, a concerted effort in the form of The Cancer Genome Atlas (TCGA) was created and serves as a critical repository of integrated molecular cancer data for researchers [26]. Despite the clear relevance of genetic aberrations in human breast cancer progression and the wide use of mouse models to study human breast cancer in vivo, relatively few mouse models of breast cancer have been characterized at the genome level, barring MMTV-PyMT [27, 28], MMTV-Neu [27], and NRL-PRL [29] models. Previous whole genome sequencing (WGS) of mouse models has revealed important human disease parallels, with sequencing of the MMTV-Neu mouse model revealing a conserved coamplification event that exists in 25% of human HER2+ breast cancers and 9% of all breast cancers. Preclinical functional studies showed the presence of this coamplification event increases migration in vitro, increases metastasis in vivo, and reduces distant metastasis free survival in humans [27]. In the MMTV-PyMT model, it was found that mutations in the phosphatase PTPRH led to increased phosphorylated EGFR levels and increased EGFR signaling [27], with CRISPR ablation and rescue experiments demonstrating that PTPRH was normally dephosphorylating EGFR [30]. In each model, genomic alterations drove key aspects of cellular signaling, altering tumor biology, with key similarities to human cancer. These studies underscore the importance of characterizing genomic alterations in mouse models of cancer and the utility of integrating gene expression and DNA sequencing of genetically engineered mouse models (GEMMs) to find clinical parallels in human cancers.

Here, we report our findings of interrogating the heterogeneity within the MMTV-Myc mouse model. We performed WGS on three common, distinct histological subtypes that arose in this model. We hypothesized genomic heterogeneity would be present across subtypes but would be conserved within each histological subtype. Importantly, each of the subtypes have numerous matched and additional tumors that have been previously analyzed for gene expression [13]. Integration of both genomic and transcriptomic data enabled the identification of active signaling pathways exploited in each histological subtype and their respective genetic origins. From this, targeted in vivo preclinical studies can be designed for the highest potential clinical translation into humans.

Methods

Whole genome sequencing

Flash frozen MMTV-Myc tumors [13] stored at −80 °C were ground using a sterile mortar and pestle under liquid nitrogen. Three randomly selected tumors each from the microacinar, squamous, and EMT MMTV-Myc histological subtypes underwent DNA extraction. DNA was extracted for each tumor using a Qiagen Genomic-tip 20/G KIT according to manufacturer specifications. Whole genome sequencing was performed at Michigan State University's Research Technology Support Facility (RTSF). DNA was sequenced at 40 × depth using Illumina HiSeq 2500 paired end 150 base pair reads after TruSeq Nano DNA library construction.

Transcriptomics

Gene expression data used in this study have previously been published [13] on MMTV-Myc and MMTV-Neu tumors and are publicly available in the Gene Expression Omnibus (GEO) under accession number GSE15904. MMTV-PyMT and MMTV-PyMT E2F knockout tumor transcriptional data have been published [31] and are available under GEO accession number GSE104397. MMTV-Neu E2F knockout transcriptional data were previously published [32] and are available under accession number GSE42533.

Whole genome sequence processing

Raw whole genome sequencing paired end fastq files first underwent initial quality control using FASTQC [33]. Quality and adapter trimming of reads were performed using Trimmomatic [34], with reads reassessed for quality afterward again using FASTQC. Reads were then aligned to the mm10 reference *Mus musculus* genome using BWA-mem [35] with option -M selected for compatibility with Picard. After alignment, read groups were added using Picard [36]. SAMtools [37] was then used to sort bam files, mark PCR duplicated sequences, and index bam files. Discordant and splitter read files were also generated and sorted using SAMtools for downstream somatic variant callers.

Somatic variant calling

Somatic mutations were called using the consensus of Mutect2 [38] (GATK suite) and VarScan [39] calls based on chromosome, position, reference base, and variant base. Variants were then annotated using SnpEff [40]. To reduce false positives, filtering included subtracting FVB-specific mutations from the mm10 C57BL/6 background as indicated in the FVB_NJ.mgp.v5.snps.dbSNP142.vcf file from the Wellcome Sanger Institute. In addition to this, HaplotypeCaller on GATK was used to call germline mutations on the FVB_NJ genome REL-1604-BAM available from the Sanger Mouse Genomes Project [41] FTP server. After converting the FVB_NJ.bam WT reference to fastq files using SAMtools, these fastq files underwent the same whole genome sequence processing as MMTV-Myc tumors. Germline variants from this wiltdtype FVB background against the mm10 reference genome were subtracted from somatic variants in each tumor.

Somatic copy number variations were determined using multiple methods; discrete copy number variations are based on the consensus of Delly [42] and Lumpy [43] calls, while whole chromosome ploidy count and segmentations were determined using CNVKIT [44]. For discrete CNVs, only those with length above 10,000 base pairs, no evidence in the WT background, and a mapping quality (MAPQ) score at 60 or above were included for analysis. Delly and Lumpy calls were combined by similar genome starting and ending positions within a 100 base pair margin of error for difference in CNV length using a custom Python script.

Inversions were taken as the consensus between Delly and Lumpy with the same restrictions and filtering steps implemented for CNVs. Translocations were called under similar approaches as CNVs and inversions. The differences are no length minimums, position differences between Lumpy and Delly calls being a maximum of 1,000 base pairs, and MAPQ scores of 50 or greater are included for analysis. To reduce false positives, these translocation calls were then merged with gene break calls made by CNVKIT by gene name. CNVKIT gene breaks were called using the "breaks" option under default options.

Every previously discussed somatic variant calling option included the FVB_NJ WT reference as the normal sample.

Pathway analysis

Previously published gene expression data from MMTV-Myc and MMTV-Neu mouse model tumors were downloaded from the GEO DataSets under the accession

number GSE15904. The gct converted file containing gene symbols as row names was used as input for ssGSEA available as a module in GenePattern [45]. Pathway analysis was conducted using the MSigDB gene set databases c1 (positional), c2 (curated), and c6 (oncogenic signature) with default settings. The resulting output data matrix with pathway enrichment scores for each sample was employed for downstream hierarchical clustering and graphical representations.

Mutation verification

Mutations observed in the whole genome sequencing analysis for APC, RARA, and KIT genes were screened and confirmed by Sanger sequencing for matched tumor samples and other tumors for a total of 15 samples (5 EMT, 5 squamous, and 5 microacinar tumors). Briefly, RNA was extracted from mammary tumors using the Qiagen RNeasy Midi KIT. cDNA was generated using the AppliedBiosystems High-Capacity cDNA Reverse Transcription KIT. Genes were amplified under PCR using the following primers: c-KIT 5′ ATAGACTCC AGCGTCTTCCG 3′ and 5′ GCTCCCAATGTCTTT CCAAAACT 3′; RARA 5′ TTGTGCATCTGAGTC CGGTT 3′ and 5′ TGGGCAAGTACACTACGAACA 3′; APC 5′ CCGCTCGTATTCAGCAGGT 3′ and 5′ CCTGCAGCCTATTCTGTGCT 3′. PCR parameters are standard parameters as listed on New England BioLabs PCR protocol (M0273) V1. PCR fragments were purified using QIAquick PCR Purification KIT or QIAquick Gel Extraction KIT. Sanger sequencing was performed by GENEWIZ/Azenta under standard premixed conditions, with one of the PCR primers used as the sequencing primer. Sequence alignment and visualization were performed with Geneious Prime-2020.0.5 software.

Circos plots

Representative mouse Circos plots for each histological subtype were generated using Circos [46] as a software package available on MSU's high-performance computing center (HPCC). All mutations, CNVs, inversions, and translocations were mapped at exact mm10 genomic coordinates of each event.

Working from the outermost region of each plot, it begins with a labeled mouse ideogram with chromosomes presented in ascending order. The next innermost rings consist of SNVs color coded to represent low (yellow), moderate (orange), or high (red) predicted impact as determined by Mutect2. Following, the next inner ring depicts copy number variations as whole integer copy number changes, with height corresponding to copy number integers of one or two, determined by CNVKIT. Copy number gains are depicted in red, while copy number losses are depicted in blue. In the final innermost circle, translocations are colored randomly to one of the two chromosomes involved in the translocation event. Inversions are colored in black. Only somatic variants that satisfy the previous requirements for somatic variant calling are included for each tumor.

Copy number and gene expression correlation

Correlation of copy number and gene expression was done on CNVKIT. After generating copy ratios and copy segments under the "batch" WGS method, discrete copy number segments were generated using the "segment" method. The "cnv_expression_correlate" method was used to generate Kendall rank correlation coefficients and Pearson correlation coefficients on discrete copy number data for all 9 tumors that underwent WGS and gene expression data for all 42 tumors that underwent transcriptomic profiling. Correlation coefficients for each gene were then mapped onto Circos plots, with significant Kendall's τ coefficients (≥ 0.3) colored blue and in the outermost ring. In the innermost ring, significant Pearson's r coefficients (≥ 0.7) are colored red. Non-significant values for both metrics were colored black. Correlation coefficients could only be generated for genes with discrete copy number changes at ± 1 of ploidy level or greater differences, so a majority of the genome will show no correlation.

Unsupervised clustering analyses

All unsupervised hierarchical clustering was performed using the clustermap function as part of the Seaborn Python library. Distance between clusters was computed using the Ward variance minimization algorithm.

PCA

Principal component analysis was performed using the scikit-learn [47] library, utilizing the StandardScaler and PCA packages. The number of principal components analyzed was 2 in every instance. Results were visualized using a custom scatter plot in matplotlib.

Copy number heatmaps

All heatmaps displaying log2 fold change of copy number segmentation data were generated using CNVKIT, both for initial copy number segmentation and copy number ratio file generation, as well as visualization.

Mutational signatures and mutational burden

Mutational burden plots were generated in matplotlib. Mutational signatures were derived using the deconstructSigs [48] R package and plotted using matplotlib.

WGS data for MMTV-Neu and MMTV-PyMT previously analyzed in the laboratory and used in mutation plots are available under the NIH SRA with BioProject

number PRJNA541842. Processing of Neu and PyMT WGS fastq reads was done according to the same methods as the MMTV-Myc samples.

Volcano plot

The volcano plot of ssGSEA c6 oncogenic signatures for the EMT subtype compared to both squamous and microacinar signatures was plotted in Python using BioinfoKIT [49]. The log-fold-change threshold is at 0.4, and the p-value threshold is set at 0.05.

Kaplan–Meier curves

Kaplan–Meier plots were made with using the Survminer R package. Publicly available and deidentified TCGA non-redundant breast cancer patient data were used to construct Kaplan–Meier curves using criteria stated for each figure.

t-SNE plot

The t-SNE diagram was created using the t-SNE implementation inside of scikit-learn. t-SNE was performed on the optimized 32 gene subset expression data of the PAM50 gene set as determined by RFECV using a support vector machine classifier. The resulting scatterplot of data is generated using matplotlib and colored according to the legend.

Supervised machine learning

The supervised machine learning soft voting classifier implemented using scikit-learn consists of logistic regression, support vector machine, random forest, XGBoost, and multi-layer perceptron classifier probabilities merged together. METABRIC gene expression data for the 32 genes with matched PAM50 subtypes were shuffled and split 70–30 into training and test data sets, respectively. To combat overfitting and training on biased data, PAM50 subtype proportions were kept the same between training and test sets. Probability predictions were averaged over 15 instantiations.

Other visualizations

Bar chart visualizations and pairplots were generated using Matplotlib, Seaborn, and Yellowbrick Python libraries.

Human data usage

Clustering was performed on human breast cancer transcriptional data matched with intrinsic subtypes from the anonymized METABRIC dataset, which is readily available through cBioPortal. Kaplan–Meier analyses were performed on non-redundant human breast cancer patients from all breast cancer studies available in cBioPortal as of time of publication that match the criteria specified in each plot.

Statistical considerations

Unless otherwise stated, statistical tests with p-value displayed are done using a two-sided Student's t-test.

Software versions

BioinfoKIT-2.1.0, Bokeh-2.4.3, BWA-mem-0.7.17, Circos-0.69.6, CNVKIT-0.9.9, Delly-0.7.8, FASTQC-0.11.7, GATK-4.1.4.1, Lumpy-0.2.13, Matplotlib-3.4.3, Mutect2-2.1, NumPy-1.20.3, Pandas-1.3.4, Panel-0.13.1, Picard-2.18.1, Python-3.9.7, SAMtools-1.9, scikit-learn-0.24.2, SciPy-1.9.0, Seaborn-0.11.2, SnpEff-4.3, ssGSEA-10.0.11, Trimmomatic-0.38, VarScan-2.4.1, and Yellowbrick-1.5.

Results

Genomic analyses reveal conserved copy number gains in microacinar tumors

Based on the diverse transcriptional and histological subtypes observed in the MMTV-Myc tumors [13], we hypothesized these phenotypes were due to a divergence in genomic changes conserved between each histological subtype. To ascertain putative conserved genomic changes within each histological subtype, short read WGS was performed on randomly selected tumors of microacinar, squamous, and EMT histological subtypes, later integrated with gene expression data obtained from matched samples as well as additional tumors from each subtype (Fig. 1). A great deal of genomic heterogeneity was observed between the histological subtypes as shown by representative Circos plots for the microacinar (Fig. 2A), squamous (Fig. 2B), and EMT (Fig. 2C) tumors. Few inversions and translocations were called across all tumors, consistent with rates of large structural rearrangements in human breast cancer [50]. Most differences in genetic aberrations between subtypes were confined to single nucleotide variants (SNVs) and copy number variants (CNVs).

Strikingly, all microacinar tumors sequenced shared the same whole chromosome amplification events on chromosomes 15 and 11 as revealed by copy number segmentation calls (Fig. 2D). Estimated total ploidy gain for chromosome 15 is 2, for a total of 4 gene copies, and ploidy gain of 1 for chromosome 11, for a total of 3 gene copies. The predicted integer copy number gains were consistent across all microacinar tumors. While many whole chromosome amplifications and deletions were observed in EMT tumors, none were consistent across the histological subtype (Additional file 26: Table S1). Interestingly, there were very few focal and whole chromosome CNVs in individual squamous tumors and none

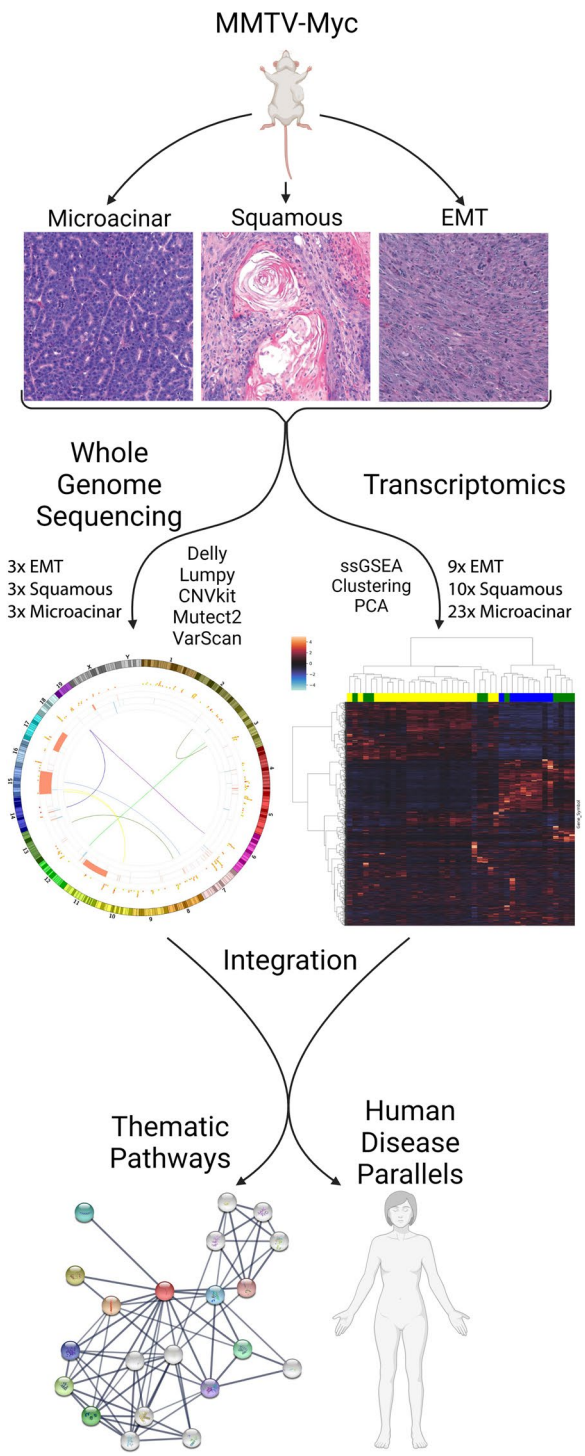

Fig. 1 Schematic of workflow. Significant histological and transcriptional heterogeneity was found in the MMTV-Myc mouse model of human breast cancer. To ascertain the genetic origins of transcriptional differences and integrate these omics data between histological subtypes, we performed short read whole genome sequencing at a depth of 40× for three tumors of each unique and predominant histological subtype: microacinar, squamous, and EMT like tumors. Somatic single nucleotide variants, copy number alterations, inversions, and translocations were profiled after alignment and processing of genomic data. Genomic somatic variants were integrated with and compared to previously obtained Affymetrix microarray gene expression and pathway analysis by single sample gene set enrichment analysis (ssGSEA). Integration of multi-omic data enables the identification of thematic pathways driving tumorigenesis and putative oncogenic drivers for each histological subtype. Importantly, these thematic pathways identified in the MMTV-Myc mouse model share similarities in human breast cancer, impactful analogous genetic events in humans co-occur with Myc amplification, and these pathways affect human breast cancer patient overall survival significantly (Created with BioRender.com)

shared across the subtype. The majority of CNVs across all tumors sequenced were broad and affected the entire chromosome rather than a discrete region within the chromosome, despite notable exceptions discussed in later figures. This is consistent with human breast cancer CNVs, which often affect a whole chromosome arm on either side of the centromere [51] except in the cases of strong selective pressure over a particular region, such as the cases of ERBB2 in breast cancer [52] or AR in prostate cancer [53].

Copy number alterations are highly correlated with gene expression

It is well established that CNVs typically correlate strongly with changes in gene expression across cancer types [54] and have been used to target cancer dependencies for therapeutic intervention [55]. Despite this, we sought to establish correlation between gene expression and CNVs in the MMTV-Myc mouse model tumors empirically. When correlating the estimated absolute integer copy number gain or loss for each gene with their matched gene expression sample and extending these results to 33 additional samples profiled by microarray, we obtained high Pearson's r and Kendall's τ correlation coefficients broadly across the genome for CNV sites (Fig. 3A). Of note, the highest and most consistent correlation for both Pearson's r and Kendall's τ occurs on chromosomes 15 and 11, suggesting the highly conserved ploidy gains across microacinar samples translates well into increased gene expression. Correlations between copy number and gene expression for EMT (Additional file 5: Fig. S1), squamous (Additional file 6: Fig. S2), and microacinar (Additional file 7: Fig. S3) specifically limited to each histological subtype are also available.

After establishing the link between CNVs and gene expression in the histological subtypes, we wanted to examine whether gene expression differences localized to human defined cytogenetic bands could stratify the

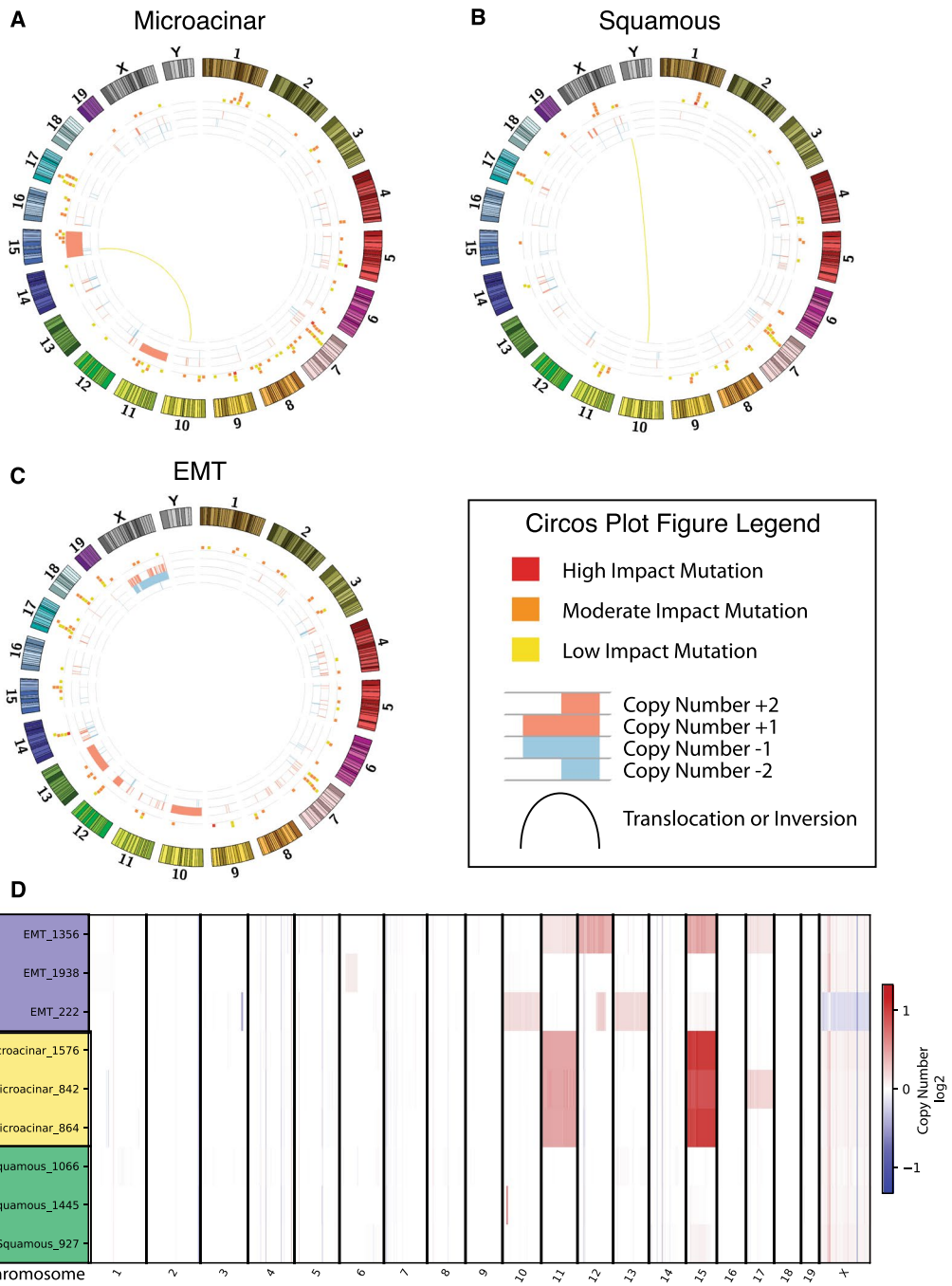

Fig. 2 Heterogeneous and conserved somatic features revealed by whole genome sequencing. Representative Circos plots are shown for (**A**) microacinar, (**B**) squamous, and (**C**) EMT histological subtypes. The outermost ring of each Circos plot depicts an ideogram for the mouse chromosomes proportionate with actual chromosome length. The next inner ring shows mutations in genes as stacked blocks at their corresponding genomic locations, color coded to their predicted impact by SnpEff [40]—yellow for low impact, orange for moderate impact, and red for high impact. The next inner ring shows discrete copy number changes as analyzed by CNVKIT; red regions indicate amplification and blue regions indicate deletions. The height of each copy number alteration corresponds to the predicted change in copy number, with the lowest level change being ± 1and showing a max copy number change of ± 2. The innermost ring reveals inversions and translocations as determined by the consensus of Delly and Lumpy somatic variant callers. Inversions are colored black, while translocations match the color of the ideogram of one of the two chromosomes involved in the translocation event. (**D**) A CNVKIT heatmap shows the $\log_2$ fold change of the estimated normalized copy number segments of each chromosome for each tumor sample relative to the wildtype reference

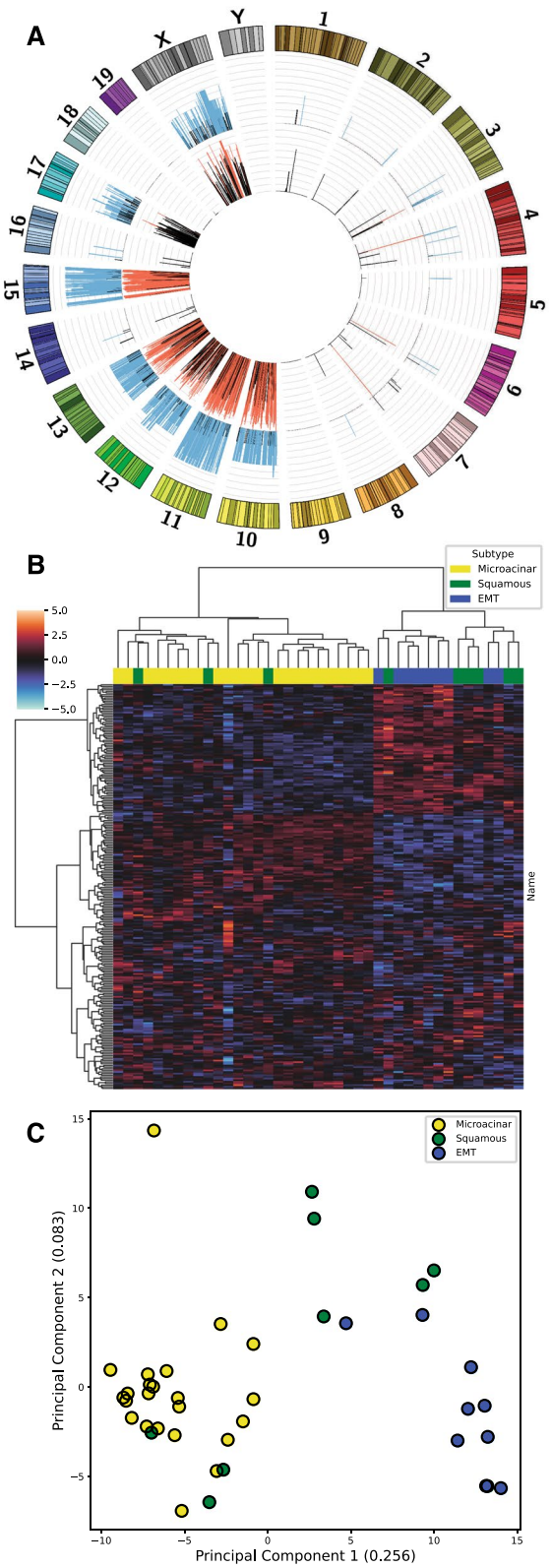

Fig. 3 Copy number changes drive gene expression changes. (**A**) Circos plot of correlation of copy number changes and gene expression across all tumor samples. Significant Kendall's rank correlation coefficient (>=0.3) shown in blue and significant Pearson correlation coefficient (>=0.7) shown in red. (**B**) Unsupervised hierarchical clustering of the MSigDB C1 positional dataset for ssGSEA values closely recapitulates the stratification of histological subtypes by gene expression clustering. (**C**) Principal component analysis (PCA) of C1 positional ssGSEA values reveals distinct clustering by histological subtype

histological subtypes. Indeed, when performing ssGSEA on gene expression data using the MSigDB C1 positional gene set, we find that unsupervised hierarchical clustering of these data post normalization largely clusters histological subtypes separately (Fig. 3B), recapitulating clustering from raw gene expression data [13]. These data suggest CNVs or other events confined to specific genomic regions are largely responsible for gene expression differences between histological subtypes. Principal component analysis (PCA) on these same C1 ssGSEA values reveals a similar clustering pattern among subtypes (Fig. 3C). Principal component 1 was able to explain over 25% of the variance in this population, which largely separated the EMT and microacinar subtypes, with squamous tumors infiltrating both clusters or being separated mostly by principal component 2.

Differential mutational landscapes between mouse models of breast cancer

Having established CNVs as being a critical factor in determining gene expression in MMTV-Myc tumors, we sought to investigate whether differing mutations within these tumors played a role in gene expression changes. After limiting gene expression data to genes predicted to have moderate or high impact mutations as determined by SnpEff, we did not see a statistically significant difference in the proportion of genes that were differentially regulated compared to the overall proportion of all genes that are differentially regulated between subtypes (data not shown). This is unsurprising as mutations typically do not affect self-gene expression, except in the case of truncating mutations [56].

Next we examined wholistic mutation patterns between each histological subtype and compared them to mutation patterns in previously sequenced MMTV-Neu and MMTV-PyMT tumors [27]. We found no large differences in overall mutational burden between MMTV-Myc subtypes, although we found trends between mouse

models (Fig. 4A). Investigating these trends, we find mutational burden did not diverge significantly between MMTV-Myc subtypes but varied considerably in the MMTV-Neu mouse model. The MMTV-PyMT model exhibited the highest overall mutational burden of mouse models analyzed, but also demonstrated considerable variability between individual tumors.

We hypothesized these differences in mutational burden could arise from separate oncogenic drivers and mutational processes within each mouse model given the extensive characterization of mutational processes in human cancers [57, 58]. To examine mutational processes in our mouse models, we utilized deconstructSigs [48] to generate different COSMIC single base substitution (SBS) signatures that take adjacent nucleotides into context and ascribes etiologies to different trinucleotide mutational patterns. We found that all MMTV-Myc tumors regardless of subtype were predominated by the homologous recombination deficient (HRR) signature (Fig. 4B). The HRR signature is strongly associated with germline and somatic mutations in BRCA1 and BRCA2 mutations in human breast cancer [57]. No mutations in BRCA1 or BRCA2 were found in any of the MMTV-Myc tumors analyzed, with these signatures possibly the result of BRCA1/BRCA2 promoter hypermethylation or other factors not analyzed. MMTV-Neu tumors were predominated by the clock-like aging signatures, while MMTV-PyMT tumors demonstrated large tobacco smoking signatures. Given that all mouse models were raised in similar controlled environments, these data suggest an unidentified endogenous C > A mutational mechanism present in MMTV-PyMT tumors.

While overall mutational burden and mutational signatures cannot parse between histological subtypes of the MMTV-Myc tumors, we reasoned that specific mutations may be associated with each subtype. Indeed, we find a small number of conserved (≥ 66% of tumors in each subtype) and impactful mutations within each subtype (Fig. 4C). Notable conserved mutations in the EMT subtype include KRAS G12D activating mutations and splice variants in SCRIB. This may be significant as others have found SCRIB cooperates with MYC for transformation and mislocalization of SCRIB within the cell, sufficient to promote cell transformation [59]. Interestingly, the squamous subtype had no discernible conserved and impactful mutations between tumors. There may be heterogeneous conservation of signaling pathways activated at the transcriptional level, though, as each squamous tumor had impactful mutations in transcription factors: zinc-finger and BTB domain containing (ZBTB) genes, zinc-finger protein (ZFP) genes, or both. For the microacinar tumors, we observed conserved missense mutations at A538E in proto-oncogene c-KIT (KIT) and at A255D for retinoic acid receptor-α (RARA). KIT is a well-established oncogene, particularly in acute myeloid leukemia (AML) [60] where mutations play a large role in pathology, and RARA is involved in embryonic development whose disruption is well established in carcinogenesis [61].

Interestingly, KIT and RARA mutations were found to be mutually inclusive in the tumors sequenced. Of the 9 tumors sequenced at the genome level, 5 were found to have both the A538E KIT and A255D RARA mutations, while the other 4 tumors contained no discernible mutations whatsoever in either KIT or RARA. To validate our WGS findings and evaluate the extent of these mutations further in other MMTV-Myc tumors, we performed Sanger sequencing on the tumors that previously underwent WGS and an additional two tumors from each histological subtype. From these data, we confirmed that these KIT and RARA mutations were mutually inclusive in a larger population of 7 of the 15 tumors that underwent Sanger sequencing. From the total populations analyzed, these mutations were present in 60% of both the microacinar and squamous subtypes, while only 20% in EMT (Additional Files 1 and 2). These data may suggest a link between KIT and RARA given their co-occurrence patterns; however, the exact functional implications of these mutations are not understood.

To gain a better understanding of the potential functional impact of the co-occurring KIT and RARA mutations, we performed multiple sequence alignment on KIT and RARA amino acid sequences from their most prevalent isoform using Clustal Omega [62]. Comparing between species, A538 KIT in the mouse maps to M535 in humans with an overall interspecies amino acid identity of 82.77% for the full protein (Additional file 3). Both mutations reside in exon 10 of their respective species, where the majority of this exon codes for the transmembrane domain in topological space. According to TCGA

(See figure on next page.)

Fig. 4 Oncogenic drivers determine mutational heterogeneity. (**A**) Total counts of overall somatic mutational burden of MMTV-Myc tumors compared to mutational burden of MMTV-Neu and MMTV-PyMT tumors as shown by bar plot. (**B**) Weights of Catalogue of Somatic Mutations in Cancer (COSMIC) mutational signatures derived from each tumor using DeconstructSigs [48] depicted by stacked bar plot. (**C**) Venn diagram of conserved mutations (≥ 66% of tumors) between histological subtypes of moderate or high impact predicted by SnpEff. Putatively impactful oncogenes with Sanger sequencing confirmed mutations are listed by their representative histological subtype in which those mutations are conserved

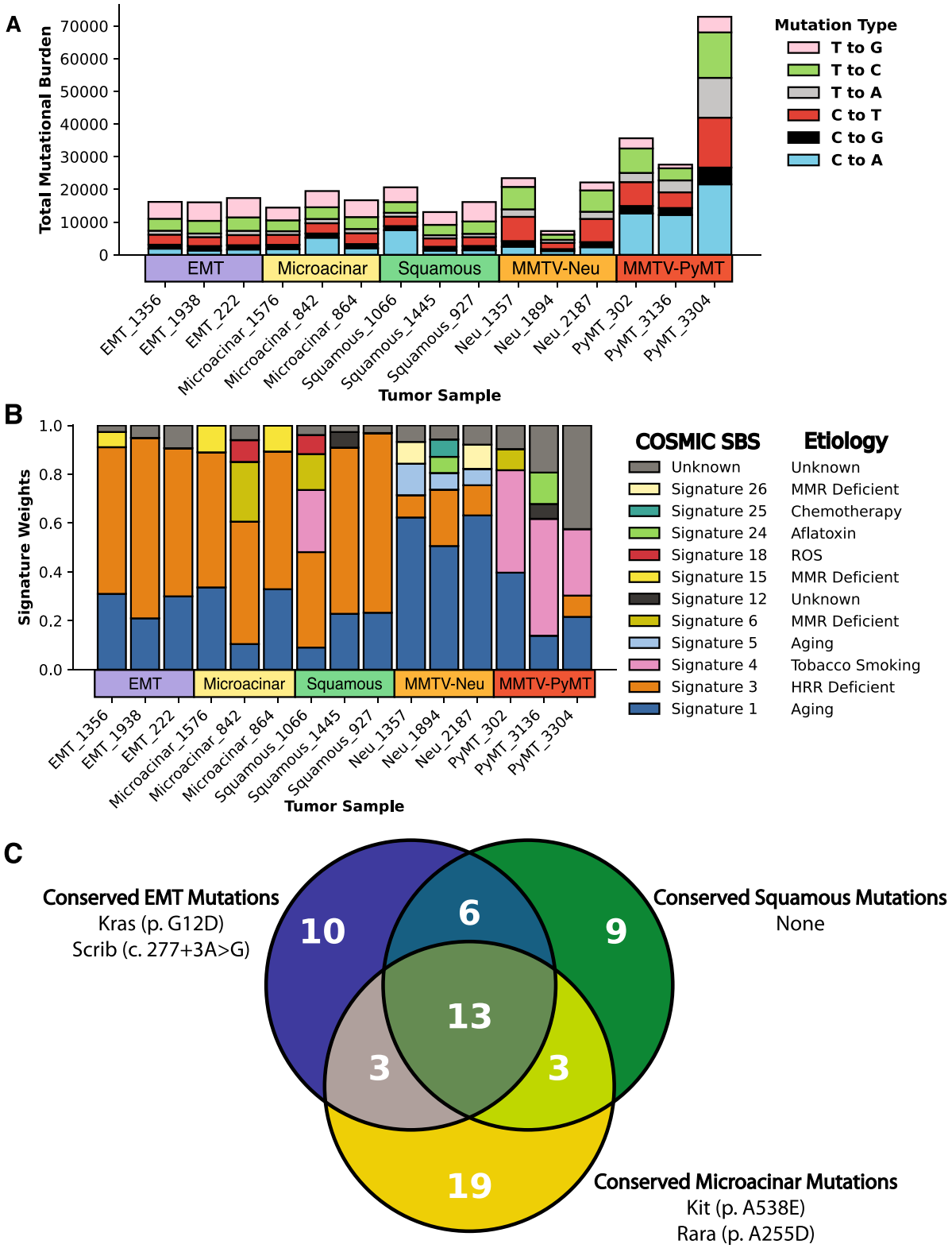

Fig. 4 (See legend on previous page.)

PanCancer Atlas data shown in cBioPortal, the KIT transmembrane domain in humans spans amino acids 525–545, suggesting that the analogous A538E mutation in the mouse occurs directly in the middle of the transmembrane domain. Additionally, while TCGA PanCancer data do not show confirmed pathogenic mutations in the transmembrane domain of KIT, the transmembrane domain resides between two confirmed mutational hotspots labeled as likely oncogenic (Additional file 8: Fig. S4). We speculate that the A538E mutation will destabilize the transmembrane domain of KIT leading to dysregulated KIT signaling. Follow up functional studies will need to be performed to ascertain the causal effects of this mutation in mice and in humans.

Similarly, A255 RARA in the mouse maps directly to A255 RARA in humans. Importantly, exon 6, where the mutations reside in both species, shares 100% identity, while the whole protein sequences share 89.98% identity (Additional file 4). Exon 6 in both species comprises the beginning portion of the hormone receptor ligand binding region of RARA. The TCGA PanCancer cohort does not contain A255D RARA mutations and most mutations are not confirmed whether they are pathogenic or not (Additional file 9: Supplementary Fig. S5). We speculate that the A255D mutation interferes with retinoic acid binding and heterodimerization with the retinoid X receptor, which thus inhibits the transcriptional activation of downstream genes that would lead to cell differentiation and cell cycle control [63, 64].

Heterogeneous activation of KRAS signaling in the EMT subtype

Previous studies on the MMTV-Myc model have shown that the EMT subtype often possessed activating KRAS mutations and increased RAS signaling [13], whereas the squamous and microacinar subtypes largely did not (Fig. 5A). This was the case for EMT tumors 222 and 1938 that underwent WGS, where both had KRAS G12D missense mutations, while the squamous tumor 1066 acquired the less transforming KRAS G13R mutation (Fig. 5B). Consequently, when comparing ssGSEA values for the C6 oncogenic signature gene sets between the three histological subtypes, EMT tumors consistently had upregulation of various KRAS signaling pathways across tissue types (Fig. 5C). Investigation of a representative KRAS signaling pathway shows the probability of KRAS activation is considerably higher in EMT than both microacinar and squamous samples (Fig. 5D).

Despite its high ssGSEA KRAS activity scores (Additional file 27: Table S2), EMT tumor 1356 possessed no KRAS mutations or mutations in genes elsewhere in the RAS pathway. Copy number segmentation data obtained pointed to a high-level amplification event on chromosome 7, encompassing FGFR2 and ATE1 (Fig. 5E). Of particular significance is the estimated copy number gain of this region. EMT tumor 222 had a similarly bounded focal amplification event over FGFR2 and ATE1 that increased both gene copy numbers by 1, but EMT tumor 1356 has an estimated copy number gain of 471 over this region and correlates well with gene expression levels of FGFR2 (Additional files 26 and 28: Tables S1 and S3). The scale of this amplification event suggests a strong selective pressure for focal amplification of FGFR2 in EMT tumor 1356. When examining the pathways of FGFR2 in the literature, FGFR2 acts directly upstream of RAS-MAPK, PI3K-AKT, and JAK-STAT signaling pathways [65] (Fig. 5F). Thus, we propose that the increase in KRAS signaling seen in EMT tumor 1356 is due to the large, focal amplification of FGFR2 and subsequent activation of RAS signaling. Stemming from the previous assessment, it is likely that the EMT histological phenotype is dependent on increased RAS signaling, ostensibly through heterogeneous genetic mechanisms.

Squamous tumors represent an intermediate phenotype between microacinar and EMT

To this point, there are strong associations between copy number amplifications specific to the microacinar subtype and heterogeneous genetic events that lead to KRAS pathway activation in the EMT subtype. However, there are no readily identifiable conserved genetic features that can explain the squamous phenotype. Despite this, squamous tumors consistently occupy a gene expression state between that of the EMT and microacinar subtypes. PCA of C2 curated gene sets (Fig. 6A) and C6 oncogenic signature gene sets (Fig. 6B) show clustering of squamous

(See figure on next page.)

Fig. 5 Heterogeneous activation of KRAS pathway in EMT histological subtype. (**A**) Proportion of each tumor histological subtype with activating mutations in KRAS in bar plot format. (**B**) Sequence variation in KRAS between all tumors sequenced as shown by a logo plot illustrates canonically activating mutations in KRAS. (**C**) A volcano plot of the ssGSEA values from the MSigDB C6 oncogenic signature gene set, showing EMT upregulated or downregulated gene sets compared to microacinar and squamous. (**D**) Violin plot of a representative KRAS pathway signature from the ssGSEA values of the C6 oncogenic signature gene set, showing distinct upregulation of KRAS signaling in the EMT subtype. (**E**) Heatmap of log2 fold change in copy number segmentation values showing high-level amplification of FGFR2 in EMT. (**F**) Canonical molecular pathway signaling reveals FGFR2 lies directly upstream of KRAS (Created with BioRender.com)

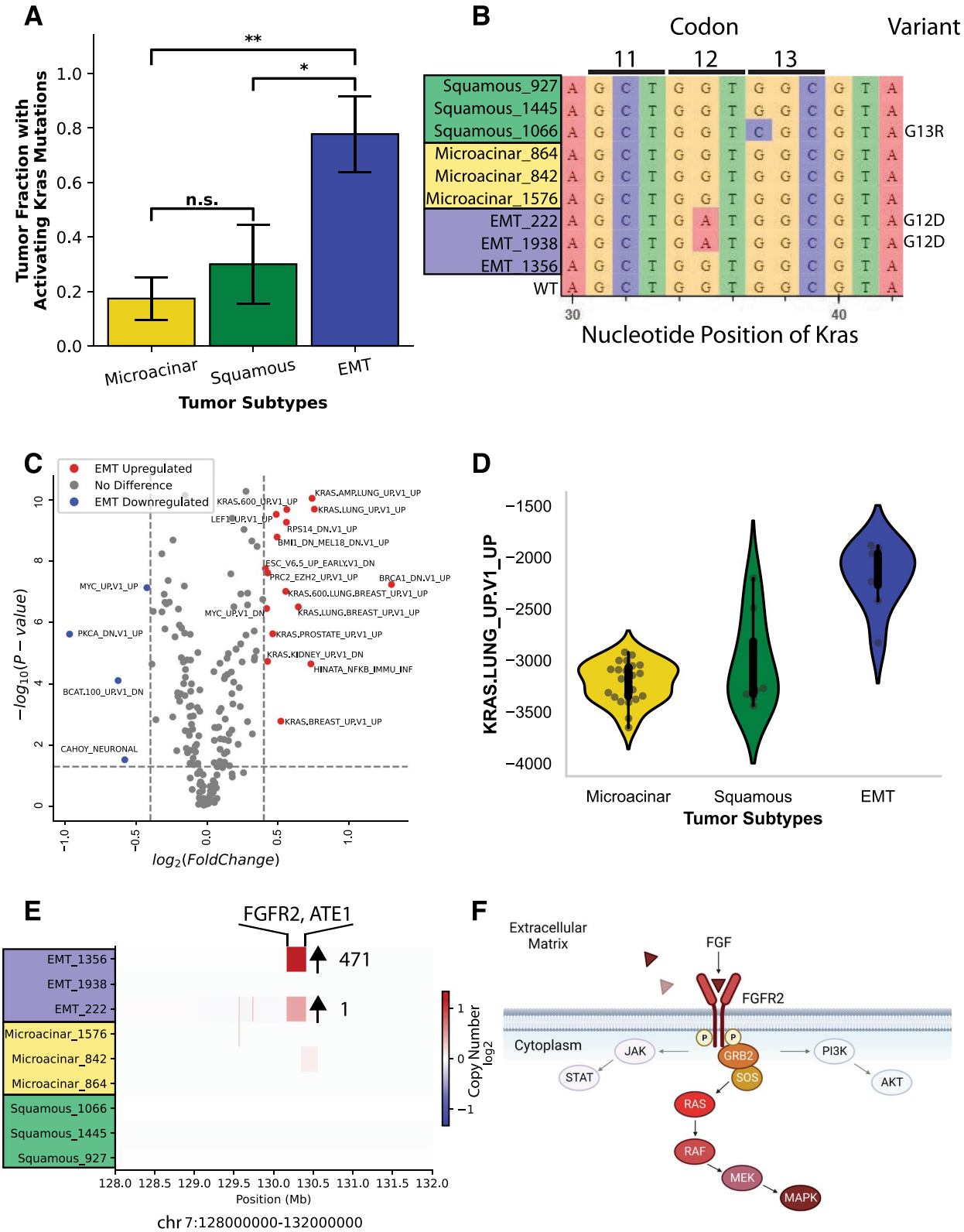

Fig. 5 (See legend on previous page.)

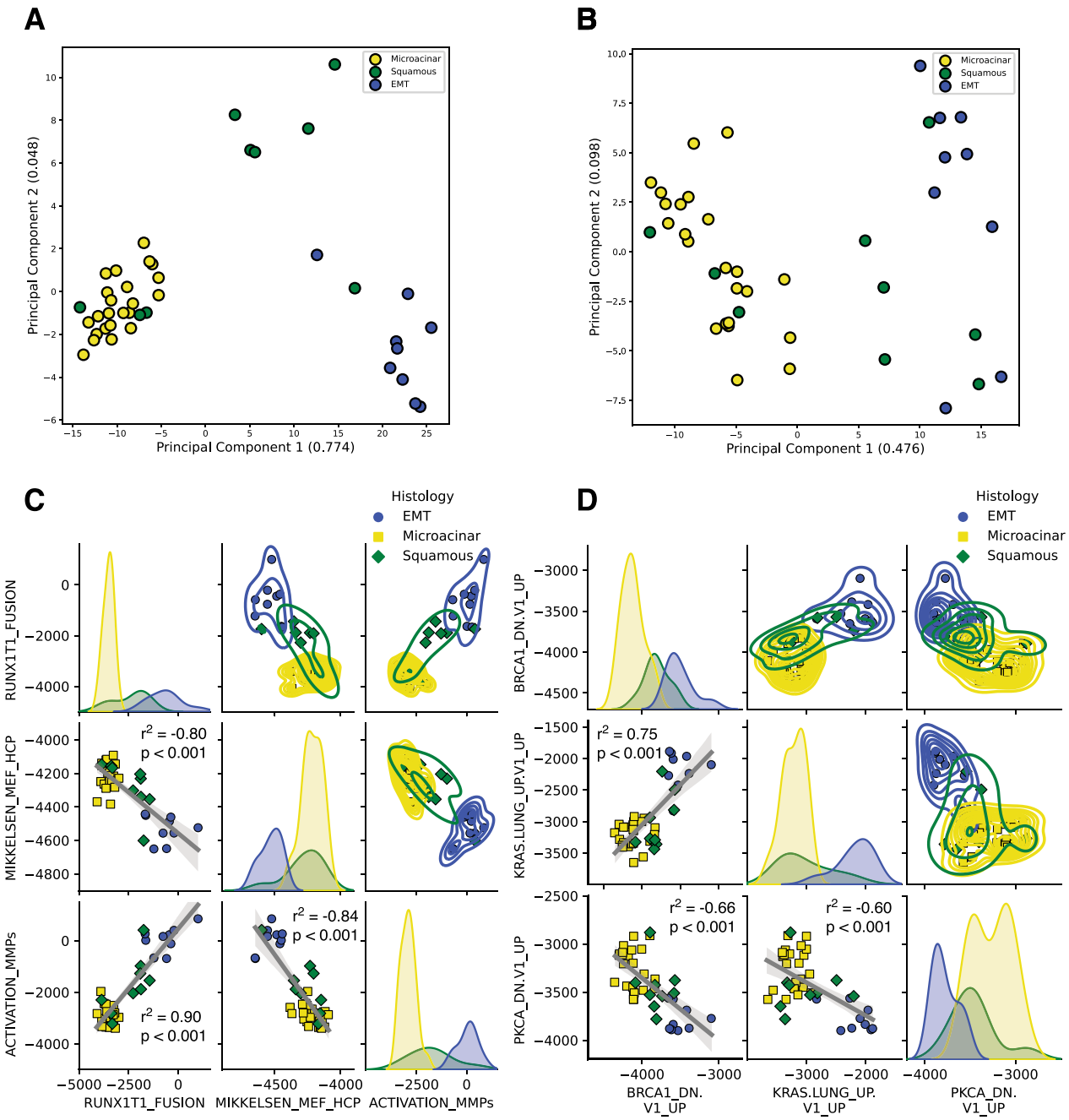

Fig. 6 Squamous represents an intermediate phenotype between microacinar and EMT. (**A**) PCA of the MSigDB C2 curated and (**B**) C6 oncogenic signature gene sets for ssGSEA values of the microacinar, squamous, and EMT tumors recapitulates C1 clustering and explains more variance in the data. (**C**) Pairwise relationship plots for representative C2 gene set and (**D**) C6 gene set ssGSEA values are shown with linear regression lines and a 95% CI. Pearson R correlation values are shown with p-values determined from a two-sided t-test

samples between microacinar and EMT, with some squamous samples invading both the microacinar and EMT clusters.

From the previous results, we sought to examine the relationship between individual pathways in each gene set that could explain these differences. To accomplish this, we chose the top three differentially expressed representative pathways from EMT and microacinar pathway signatures for C2 (Fig. 6C) and C6 (Fig. 6D) MSigDB gene sets. While there were no statistically significant correlations within each histological subtype (data not shown), likely due to limited numbers of samples,

pairwise correlations of pathway signatures across all three subtypes showed both significant positive and negative correlations between ssGSEA pathway activities. Beyond this, top differentially expressed pathway activities appear to lie on a continuum, with squamous tumors routinely spanning between and infiltrating the microacinar and EMT clusters. Pairwise relationships for the C1 positional gene sets shows similar patterns (Additional file 10: Fig. S6). These data are consistent with the squamous histology occupying an intermediate phenotype between that of EMT and microacinar subtypes.

Integrated mouse data stratifies human breast cancer subtypes and yields clinical insights

It is clear that the MMTV-Myc mouse model produces primary mammary tumors that are heterogeneous in histology, gene expression, metastatic variance [13], and now somatic genomic perturbations that can explain many of the transcriptional differences seen in this model. However, the significance of these events and translational potential to humans is not immediately obvious.

To evaluate whether the events seen in MMTV-Myc mouse model could have predictive power in clinical outcomes, we utilized an integrative approach, combining gene expression data and somatic genetic events to examine how they can parse human breast cancer subtypes. To this end, we took all genes that were differentially expressed between MMTV-Myc histological subtypes that were also present in conserved copy number gain or loss events to obtain the integrative gene set (Additional file 29: Table S4). Subsequently, we performed unsupervised hierarchical clustering on human gene expression from the METABRIC breast cancer dataset [66–68] limited to the integrative mouse gene set (Fig. 7A). We found distinct clusters emerged that well represented the intrinsic subtypes of breast cancer. Importantly, clustering from the same number of genes used but instead randomly selected failed to resolve intrinsic breast cancer subtype clusters to the same degree (Additional file 11–20: Figs. S7–S16). This suggests that the integrated mouse gene set represents a diverse set of informative genes across all intrinsic subtypes of human breast cancer that can effectively differentiate between subtypes rather than contributing noise. Clustering by the PAM50 gene set showed improved subtype clustering overall relative to the mouse integrative gene set but did not resolve the luminal A and luminal B subtypes to the same extent (Fig. 7B). It is important to note that the PAM50 gene set is curated specifically to be able to differentiate between human intrinsic breast cancer subtypes, so seeing improved performance relative to the integrative mouse gene set is expected.

It is clear that genetic events in the MMTV-Myc mouse model and their resulting gene expression differences can resolve human breast cancer intrinsic subtypes. However, it is unclear to what extent these genetic events in the MMTV-Myc mouse model represent genetic events occurring in human breast cancer. To address this, we assayed publicly available TCGA datasets on breast cancer available through cBioPortal. We assessed prevalence of genetic events in human breast cancer first identified to be conserved both in genomic alteration status and differential gene expression in the MMTV-Myc mouse model. Subsequently, we examined the effects of these genes on human breast cancer overall survival clinical outcomes through Kaplan–Meier analysis.

All of the genetic events examined were found to be co-occurring with MYC amplification in human breast cancer (Additional file 30: Table S5), especially supporting the use of the MMTV-Myc mouse model in studying MYC-driven human breast cancers. However, because of the co-occurring nature of these events and the limited number of patient samples with genetic events and matched clinical outcomes, statistical significance is difficult to achieve in mutually exclusive populations of MYC amplification and identified genetic events. MYC amplification and overexpression are well described in TNBC and basal-like subtypes of breast cancer [1, 69], known as the deadliest subtypes of breast cancer currently [70, 71], which must be accounted for in survival analysis. To remedy this, we look at relative differences in overall survival from MYC amplification compared to analogous genetic events in humans identified in the mouse model.

Kaplan–Meier analysis on breast cancer patients revealed MYC amplification was present in 18.2% of patients surveyed, with median overall survival at 135.2 months (95% CI: 113.7–148.1) compared to the unaltered population with median overall survival at 171.3 months (95% CI: 161.2–182.9) (Fig. 7C, Additional files 31 and 32: Tables S6 and S7). When comparing the patient population of those harboring KRAS activating mutations or amplifications, which account for 2.2% of all breast cancer cases, we find a marked decrease in overall survival compared to the unaltered cohort (Fig. 7D) with median overall survival at 77.7-months (95% CI: 61.8–146.4) for the KRAS altered population compared to median overall survival of 164.3 months (95% CI: 154.3–173.0) for the unaltered cohort (Additional files 33 and 34: Tables S8 and S9). While it cannot be ruled out that MYC amplification co-occurrence in the KRAS altered cohort reduces overall survival, the KRAS altered population maintains a 57.5-month median overall survival deficit to MYC amplification alone. Thus, MYC amplification may only have a modest additive effect to KRAS amplification or

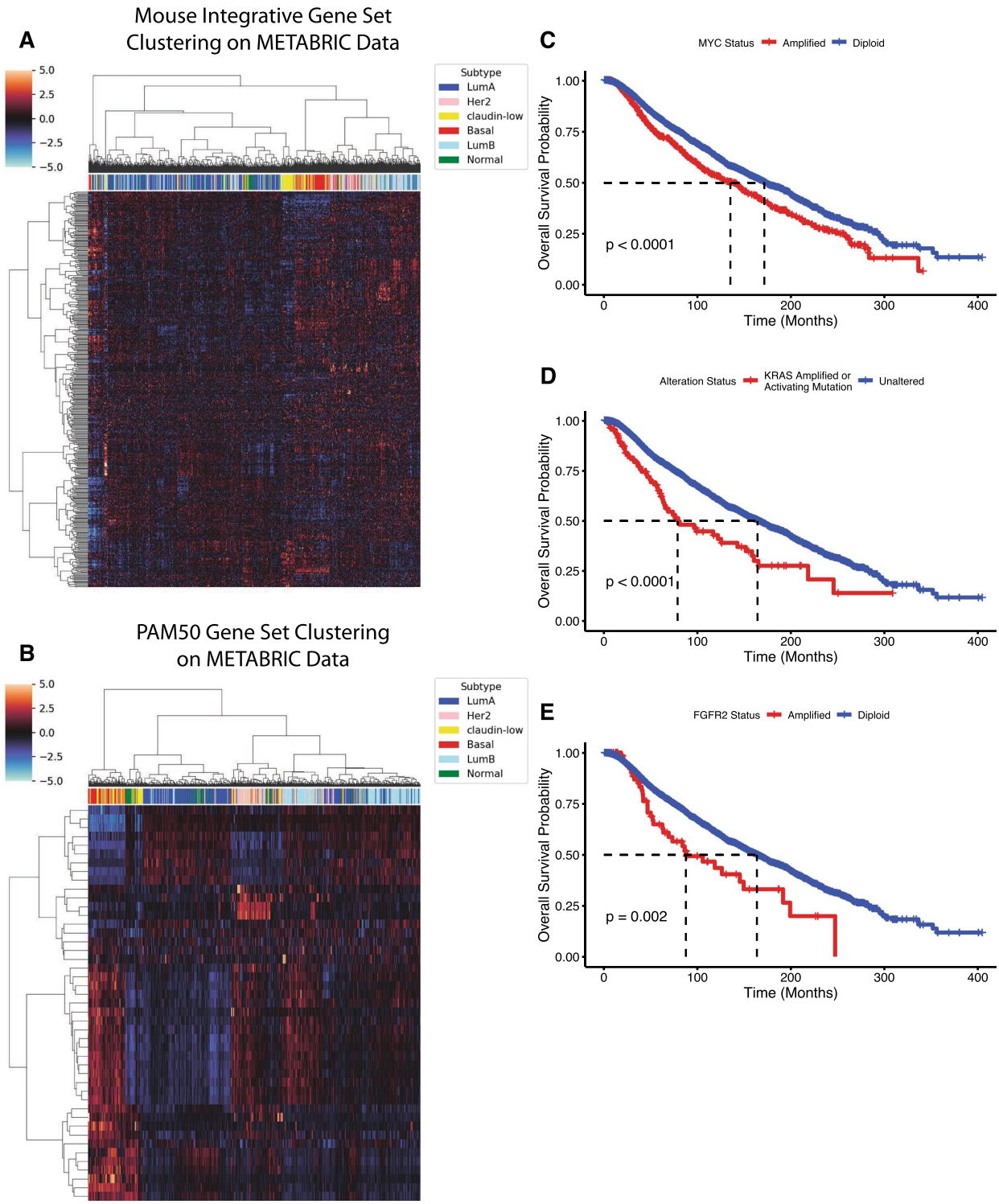

Fig. 7 Mouse model genetic and transcriptional events inform human clinical outcomes. (**A**) Unsupervised hierarchical clustering of METABRIC gene expression values by a list of 453 homologous genes that are in a conserved amplification/deletion event and are differentially expressed between MMTV-Myc histological subtypes. (**B**) Unsupervised hierarchical clustering of METABRIC gene expression values by the PAM50 gene set. (**C**) Overall survival (OS) Kaplan–Meier (KM) analysis of non-redundant TCGA breast cancer patients, accessed through cBioPortal, stratified by Myc amplification status. (**D**) OS KM curve of non-redundant TCGA breast cancer patients stratified by KRAS alteration status. (**E**) OS KM curve of non-redundant TCGA breast cancer patients stratified by FGFR2 amplification status

activating mutations in this cohort. It is worth noting that the unaltered population in the KRAS cohort has reduced overall survival compared to the MYC unaltered cohort, suggesting most patients with MYC amplification are largely shunted to the unaltered group.

Similarly, patients with focal FGFR2 amplifications, accounting for 1.5% of all breast cancer cohort patients, also exhibit a marked decrease in overall survival compared to the unaltered cohort (Fig. 7E) with median overall survival at 87.7 months (95% CI: 62.4–191.0), contrasting with the unaltered cohort at 163.5 months (95% CI: 154.0–172.9) (Additional files 35 and 36: Tables S10 and S11). Again, patients with FGFR2 amplification fare significantly worse than those with MYC amplification, standing at a difference of 47.5-month median overall survival difference. While KRAS and FGFR2 alterations are infrequent in breast cancer, these alterations may be extremely significant in the prognosis and treatment of their disease.

Altogether, these analyses reveal conserved genetic events between both human Myc-driven breast cancer and the MMTV-Myc mouse model of breast cancer. Importantly, these somatic genetic events are associated with severe drops in overall survival in human breast cancer patients in large excess of what Myc amplification causes.

Machine learning classifier predicts MMTV-Myc histological subtypes correspond to different human breast cancer intrinsic subtypes

Thus far, we have shown that there exist tightly linked transcriptional and genomic perturbations in the MMTV-Myc model, which are heterogeneous across its histological subtypes with clinical implications in humans. However, whether these histological subtypes are representative of different human breast cancer intrinsic subtypes is unclear. Unsupervised clustering by the integrative mouse gene set shows modest ability to parse human intrinsic breast cancer subtypes, but it falls short of being able to discriminate between features (genes) that can best exemplify a given class (intrinsic subtype). To this end, we employed a machine learning model classifier trained on human METABRIC gene expression data to predict which human breast cancer intrinsic subtype each MMTV-Myc histological subtype best represents.

To accomplish this, we first combined the raw METABRIC microarray gene expression data with the MMTV-Myc microarray data (GSE15904), along with additional mouse microarray cohorts for variations of the MMTV-Neu mouse model (GSE42533) and the MMTV-PyMT mouse model (GSE104397) for comparison. PCA of normalized data revealed strong batch effects between datasets that would distort causal biological interpretation (Additional file 21: Fig. S17). In removing batch effects, we employed the parametric empirical Bayes shrinkage adjustment available from ComBat [72], effectively eliminating non-biological differences between datasets (Additional file 22: Fig. S18).

We then sought to narrow down the number of features in this dataset to avoid overfitting the model and make it more generalizable to new patient data for intrinsic subtype prediction. The prediction analysis of microarray 50 (PAM50) is a well-established scoring metric of gene expression data for 50 genes to stratify breast cancer patients into intrinsic subtypes and offer clinical prognoses [73]. From here, we employed recursive feature elimination with cross-validation (RFECV) with a support vector machine (SVM) radial basis function (RBF) kernel estimator to determine the optimal features in the dataset. Surprisingly, 32 specific genes (Additional file 37: Table S12) from the PAM50 subset gave higher cross-validation scores than utilizing all 50 genes (Additional file 23: Fig. S19). In testing various estimators for the machine learning classifier model limited to this 32 gene subset, we found a great deal of variability between models in hard classification predictions between model instantiations. To alleviate this, we utilized a soft voting classifier composed of logistic regression, SVM with RBF kernel, random forest, XGBoost, and multi-layer perceptron estimators averaged over 15 k-fold instantiations. Pooling multiple estimators together was done to reduce bias from any one particular estimator and to increase overall accuracy. Subsequently, the highest average probability of a given class will decide which intrinsic subtype a mouse tumor belongs to.

The motivation for developing a supervised machine learning model classifier is evident after examining the distribution of the combined human and mouse gene expression dataset through a scatterplot of t-distributed stochastic neighbor embedding [74] (t-SNE) data in two-dimensional space (Fig. 8A). There are distinct clusters of human breast cancer samples forming the intrinsic subtypes except for claudin-low, which supports the notion that claudin-low breast cancers represent an additional complex phenotype rather than an intrinsic subtype of breast cancer [75]. The mouse samples are well dispersed throughout the plot even within the same mouse model, pointing to substantial gene expression heterogeneity within each model. However, problems arise in that it is not immediately obvious which cluster each mouse sample belongs to. When using t-SNE, there is also necessarily a loss of information even when performing non-linear dimensionality reduction, in this case from 32 to 2 dimensions. This is nothing to say of the tendency for gradient descent algorithms such as t-SNE to become

stuck in local optima. Additionally, other unsupervised methods such as hierarchical clustering equally weight all features, which decreases accuracy of the model. For these reasons and more, there was a clear need to develop a supervised machine learning classifier using the 32 gene dataset.

Accuracy scores and *F*1 scores are often used to evaluate the efficacy of supervised machine learning models. However, it has been shown that these metrics can be overinflated [76], and so we also report the more robust Matthew's correlation coefficient (MMC) metric, which proportionally accounts for true positives, true negatives, false positives, and false negatives. For the soft voting classifier used to predict mouse-to-human subtypes, we obtained an average accuracy score of 80.0%, a weighted *F*1 score of 80.0%, and an MCC score of 74.0% after 15 k-fold stratified and shuffled cross-validation using a 70% train and 30% test split (Fig. 8B). All metrics used to evaluate the machine learning model are in high agreement and show the model is effective at predicting human breast cancer intrinsic subtypes. To better visualize these predictions on the test data, we constructed a confusion matrix showing true classes on the vertical axis and predicted class on the horizontal axis (Fig. 8C). Many of the most confused classes correspond well to the t-SNE visualization where clusters overlap, including the overlap of luminal A and luminal B classes, overlap of luminal A and normal classes, and the overlap of claudin-low and basal classes. All metrics combined show that the initial METABRIC gene expression data limited to the 32 most predictive genes shared with the mouse microarray data has significant power in discriminating between intrinsic subtypes of breast cancer using the soft voting classifier.

When applying the classification model to mouse gene expression data, we obtain heterogeneous subtype predictions across the various MMTV-Myc-, MMTV-Neu-, and MMTV-PyMT-based mouse models (Fig. 8D). Our initial hypothesis postulated that the MMTV-Myc model would enrich for claudin-low and basal-like tumors, similar to how human breast cancer patients with amplified MYC are preferentially basal-like or claudin-low. Indeed, for MMTV-Myc tumors overall, 30% are predicted as claudin-low, 24% are luminal A, 16% are luminal B, 11% are basal, 11% are HER2+, and 8% are normal-like. Large differences in intrinsic subtype proportions are evident across all mouse models tested in comparison with human intrinsic breast cancer subtype proportions, with METABRIC intrinsic subtype proportions at 35% luminal A, 24% luminal B, 11% HER2+, 11% claudin-low, 11% basal, and 8% normal-like. The claudin-low subtype is especially enriched across all mouse models examined.

Though, ratios of each intrinsic subtype vary considerably between histological subtypes for the MMTV-Myc model (Fig. 8E). Adenocarcinoma tumors are primarily categorized as claudin-low at 57%, with luminal B and basal subtypes trailing at 15% and 14%, respectively. Papillary tumors maintain roughly similar proportions of basal (17%), claudin-low (25%), HER2+ (19%), and luminal B (24%) subtypes. EMT tumors are overwhelmingly claudin-low (44%) and luminal B (33%). Microacinar tumors show the most enrichment for the HER2+ subtype at 21% and tied for most basal enriched with papillary at 17%. Squamous tumors are considerably variable, with 33% classified as luminal A, 22% as claudin-low, and 21% as normal-like, corroborating previous pathway signatures showing squamous tumors are not confined to a specific localized cluster.

These results have the potential to inform mouse model use when investigating different subtypes of breast cancer or examining breast cancer heterogeneity more generally. For instance, the MMTV-Neu mouse model is often used as a model of HER2+ breast cancer, but most MMTV-Neu tumors show gene expression similar to luminal A and claudin-low human tumors. While other groups have predicted MMTV-PyMT tumors to correspond to the luminal B subtype [19], we predict only 13% of PyMT based mouse models fall into this category, with PyMT tumors overall being quite heterogeneous.

In summary, we have created an accurate supervised machine learning classification model that can stratify human breast cancer intrinsic subtypes. When applied

(See figure on next page.)

Fig. 8 MMTV-Myc histological subtypes are representative of different human breast cancer intrinsic subtypes. (**A**) t-distributed stochastic neighbor embedding (t-SNE) performed on human METABRIC and mouse model gene expression samples using the 32 homologous gene subset of PAM50 as determined by recursive feature elimination with cross-validation. (**B**) Normalized scoring metrics for the soft voting classifier including accuracy for measuring true positives, a weighted F1 scoring metric for balancing precision and recall, and a Matthews correlation coefficient (MCC) metric for taking into account false positives and false negatives even in the case of unbalanced classes. (**C**) A confusion matrix where true positives lie along the diagonal from top left to bottom right and false values occupy all other boxes. (**D**) Bar chart of proportional probabilities of each model representing human intrinsic breast cancer subtypes as determined by the soft voting classifier. Human intrinsic subtype proportion was determined directly from proportions of METABRIC breast cancer patients subtyped. (**E**) Bar chart of MMTV-Myc histological subtypes and proportional probabilities of each subtype corresponding to different human breast cancer intrinsic subtypes determined by the soft voting classifier

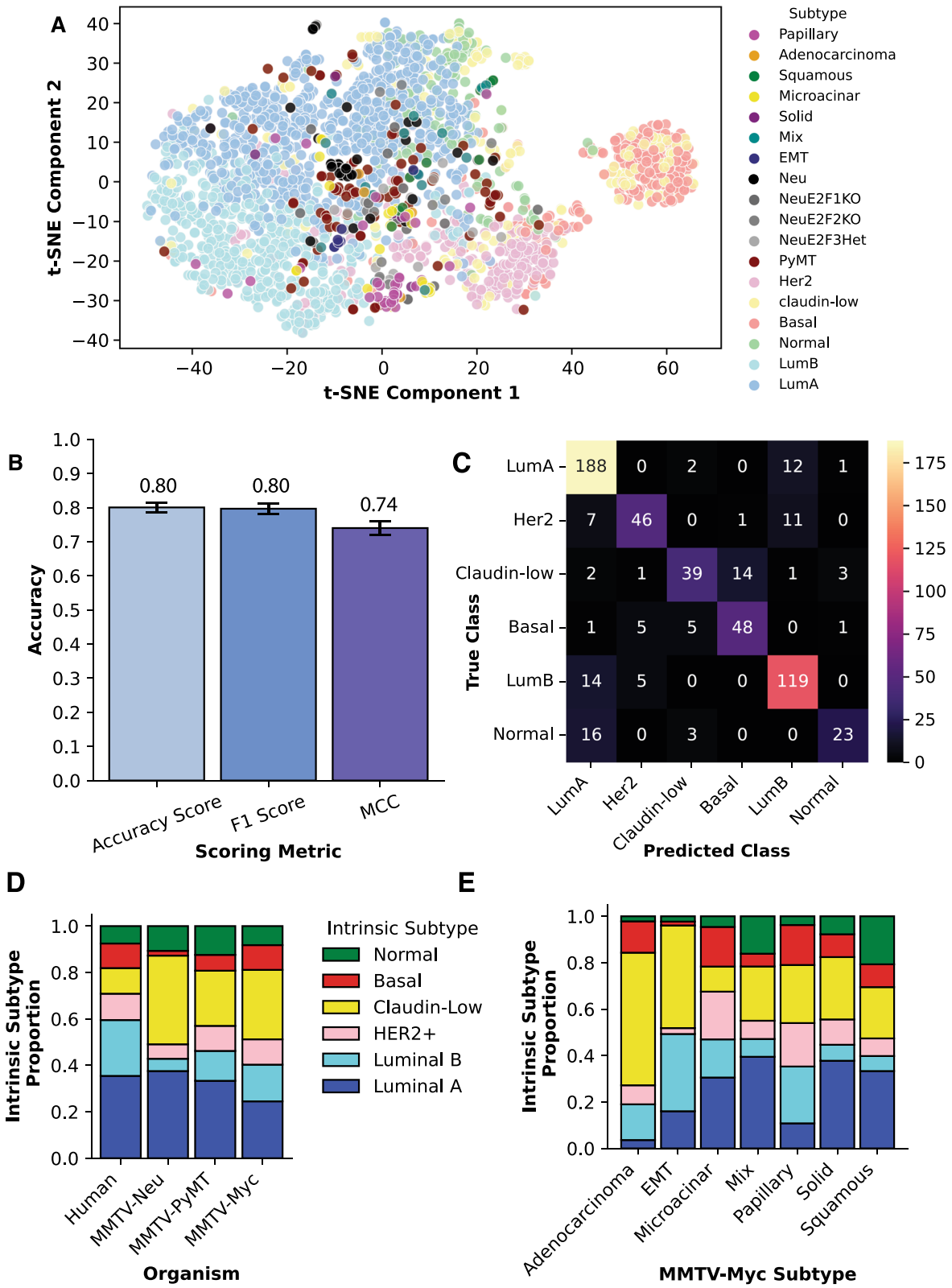

Fig. 8 (See legend on previous page.)

to batch effect corrected mouse transcriptional data, we observe diverse intrinsic subtype profiles assigned to different histological subtypes from the MMTV-Myc mouse model.

Discussion

The utility of mouse models of breast cancer has progressed from overexpression of a driving oncogene to questions as to whether they accurately mimic the heterogeneity and progression of human breast cancer. Here, we have described the utility of the MMTV-Myc GEMM in recapitulating the histological, transcriptional, and genomic heterogeneity of human breast cancer.

Conserved somatic genetic changes across MMTV-Myc histological subtypes are associated with negative effects on overall survival when applied to human breast cancer patients. These somatic events have largely been overlooked previously due to their low prevalence in human breast cancer and resulting lack of statistically significant differences in clinical outcomes. Given that MYC is frequently amplified in basal-like and TNBCs [1], and the current lack of targeted therapies in TNBCs [77], it is likely there is no consistent driver of oncogenesis across all TNBC patients. Instead, we hypothesize there may be low prevalence oncogenic drivers that lead to similar transcriptional profiles ultimately, but each patient's tumor maintains different genetic drivers. An example that arises in this paper is that of activating KRAS mutations or significant ploidy gain of FGFR2. Activation of either proto-oncogene will result in increased mitogen-activated protein kinase (MAPK) signaling, but treating the root oncogenic driver will require different therapeutic strategies. The most prevalent KRAS mutations in breast cancer are G12C/D/V/A mutations, with some of these G12C mutant patients potentially benefitting from treatment with the recently developed sotorasib [78] or adagrasib [79] therapies. However, patients with amplified FGFR2 would not respond to these therapies and instead would more likely benefit from highly selective FGFR inhibitors such as AZD4547. A recent clinical study found that treating endocrine therapy-resistant breast cancer patients with AZD4547 achieved partial response in some patients, with differentially expressed genes involved in FGFR signaling able to distinguish between responders and non-responders [80]. Although FGFR1 is amplified more often in breast cancer, both FGFR1 and FGFR2 amplified and overexpressing breast cancers could likewise benefit from AZD4547 treatment.

OBSCN was identified with conserved mutations across all histological subtypes, with OBSCN mutations correlating poorly with overall survival in humans (Additional file 24: Fig. S20, Additional files 38 and 39: Tables S13 and S14). However, OBSCN is a large gene where mutations often co-occur with other large genes, such as TTN, MUC family genes, and FAT family genes (Additional file 40: Table S15). Therefore, mutations in OBSCN are likely biomarkers of hypermutational burden rather than OBSCN playing a tumor suppressive or oncogenic role. This could still be useful information, as hypermutant individuals are more likely to respond to immune checkpoint inhibitors [81].

Aside from direct clinical implications in humans, these results reveal human intrinsic subtype analogs in the MMTV-Myc, MMTV-Neu, and MMTV-PyMT mouse model histological subtypes. Others have classified mouse models of breast cancer and ascribed them to different human intrinsic subtypes of breast cancer previously in unsupervised methods [18, 19]. However, these analyses are not weighted for genes that are able to discriminate between intrinsic subtypes; they may be biased toward noise in the dataset rather than predictive signals and do not produce metrics for scoring accuracy of the model. To rectify this, we created an accurate machine learning classification model trained on human gene expression data and applied to batch effect corrected gene expression data of different mouse models of breast cancer. From these results, we see there is substantial heterogeneity within the histological subtypes of the MMTV-Myc model.

However, the proportions in MMTV-Myc tumors that match human intrinsic subtypes are skewed relative to their occurrence in humans. We see a general decrease in luminal tumors and an increase of claudin-low and normal-like tumors. While this is true overall for the MMTV-Myc mouse model, it is highly dependent on the histology of the tumor. For example, adenocarcinoma tumors are 57% claudin-low, while microacinar tumors are 11% claudin-low. EMT tumors are largely mapped to claudin-low and luminal B intrinsic subtypes. The EMT subtype maintains great variability in CNVs, similar to that of human luminal B breast cancer [82], but the gene expression signatures of EMT tumors overlap with the canonical signatures associated with human claudin-low breast cancer: high expression of markers for cytotoxic T-cell and natural killer (NK) cell infiltration (Granzymes C, D, E, F, and G), high expression of dormancy markers (NR2F1), and low expression of cell-adhesion proteins (CLDN2, GJB1, CEACAM1) [83] (Additional file 25: Fig. S21). These observations may be useful to cancer researchers when selecting a mouse model for studying a specific subtype of breast cancer, particularly the adenocarcinoma or EMT histological subtypes as there are no established spontaneous models of breast cancer that exclusively mimic claudin-low breast cancer [19]. These data suggest that the adenocarcinoma or EMT histological subtypes from the MMTV-Myc GEMM could be a

reliable immunocompetent model for claudin-low breast cancer.

It is clear that the copy number changes identified among the MMTV-Myc tumors sequenced, particularly the conserved ploidy gains on chromosomes 11 and 15 in the microacinar tumors, correlate with gene expression changes. ERBB2 and related genes lie on chromosome 11, which may explain the propensity of microacinar tumors to be HER2-like than other histological subtypes. It is likely these CNVs are causal in driving gene expression changes between histological subtypes, although we cannot confirm that with these limited data. CRISPR knockout followed by gene addback experiments could be used to validate these findings. The highly conserved copy number gains seen in the microacinar tumors suggest a strong selective pressure for amplification and overexpression of genes in these regions. Upon examining the human homologs and their syntenic chromosomal regions for the integrated mouse gene set, we find the two largest high synteny regions are the entirety of chromosome 17 following from the mouse chromosome 11 amplification and the long arm of chromosome 8 (8q) following from the mouse chromosome 15 amplification (Additional file 41: Table S16). It is interesting to note that human chromosome 8q contains the MYC locus and chromosome 17q contains the ERBB2 locus, with amplification of these two loci highly correlated in human breast cancer (Additional file 42: Table S17). Regions 8q and 17q are among the most frequently amplified regions in human breast cancer [84, 85]. However, given the small sample size of microacinar tumors used and experimental setup, it is impossible to determine whether either of these amplifications has causal implications for the other. It should also be noted that chromosome 17p is often deleted in many human breast cancers while the entirety of chromosome 17 in humans is able to be mapped to mouse chromosome 11. Mouse chromosomes are telocentric, and given that these frequent large amplifications lie on either side of the centromere in humans, it is possible that amplification of the analogous genes from region 17q is of higher selective consequence than deletion of the analogous genes from region 17p. Comparative genomics between the MMTV-Myc histological subtypes and MYC-driven human breast cancers may be an important area of future study.

Conclusions

A significant hurdle in the study of breast cancer in vivo has been the limitations of mouse models recapitulating the heterogeneity found in human breast cancer. Here we report that the MMTV-Myc GEMM recapitulates the histological, transcriptional, and genomic heterogeneity found in human breast cancer, with important clinical parallels identified. We find different MMTV-Myc histological subtypes preferentially represent different human intrinsic breast cancer subtypes, further solidifying the MMTV-Myc model as an appropriate in vivo method for examining the multi-faceted aspects of human breast cancer heterogeneity even down to the gene expression level.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13058-023-01723-3.

Additional file 1: Aligned sanger sequencing results of 5 microacinar, 5 squamous, and 5 EMT tumors over KIT. Sequencing shows a conserved C to A mutation at consensus sequence position 100 and present in tumors 525-1, 598-1, 864-1, 1052-1, 1356-2, 1445-1, and 1576-1. An electropherogram of some sequences is included.

Additional file 2: Aligned sanger sequencing results of the same 5 microacinar, 5 squamous, and 5 EMT tumors over RARα. Shows a conserved C to A mutation present at position 132 in the electropherogram. These mutations occur in the same tumors as the KIT C to A mutations (525-1, 598-1, 864-1, 1052-1, 1356-2, 1445-1, and 1576-1).

Additional file 3: Human (NP_000213.1) and mouse (NP_001116205.1) KIT protein sequence alignment using Clustal Omega 1.2.4. The percent identity matrix calculated by Clustal 2.1 is shown at the bottom. The percent identity between human and mouse KIT protein is 82.77%.

Additional file 4: Human (NP_000955.1) and mouse (NP_001169999.1) RARA protein sequence alignment using Clustal Omega 1.2.4. The percent identity matrix calculated by Clustal 2.1 is shown at the bottom. The percent identity matrix between human and mouse RARA protein is 89.98%.

Additional file 5: Fig. S1. Circos plot of Pearson correlation coefficient values for integer copy number and gene expression float value for the EMT histological subtype alone. Red indicates that the correlation value is above 0.7, while being colored black means the correlation value is below 0.7.

Additional file 6: Fig. S2. Circos plot of Pearson correlation coefficient values for integer copy number and gene expression float value for the squamous histological subtype alone. Red indicates that the correlation value is above 0.7, while being colored black means the correlation value is below 0.7.

Additional file 7: Fig. S3. Circos plot of Pearson correlation coefficient values for integer copy number and gene expression float value for the microacinar histological subtype alone. Red indicates that the correlation value is above 0.7, while being colored black means the correlation value is below 0.7.

Additional file 8: Fig. S4. Lollipop plot for human KIT (isoform 1) protein obtained from cBioPortal. Patient samples are from the TCGA PanCancer Atlas. The lollipops represent different mutations, with light green representing missense mutations of unknown pathology. Dark green mutations are confirmed pathologic missense mutations. Gray mutations represent truncating mutations of unknown pathology. Beige mutations represent splice variants of unknown pathology. Brown mutations represent inframe deletions or insertions of unknown pathology. Dark brown mutations represent confirmed pathologic inframe deletions or insertions. The annotation tracks below the plot correspond to the labels designated on the left. Each dot on the annotation track maps back to that location on the protein.

Additional file 9: Fig. S5. Lollipop plot for human RARA (isoform 1) protein obtained from cBioPortal. Patient samples are from the TCGA PanCancer Atlas. The lollipops represent different mutations, with light green representing missense mutations of unknown pathology. Dark green mutations are confirmed pathologic missense mutations. Gray

mutations represent truncating mutations of unknown pathology. Beige mutations represent splice variants of unknown pathology. Brown mutations represent inframe deletions or insertions of unknown pathology. The annotation tracks below the plot correspond to the labels designated on the left. Each dot on the annotation track maps back to that location on the protein.

Additional file 10: Fig. S6. Pairwise correlation plots of MSigDB C1 positional gene sets for EMT, microacinar, and squamous tumors using gene sets (from top to bottom and from left to right): chr22q11, chr8p21, chr17q22, chr11q23, chr8p12, and chr3q12. Generated using the Seabron pairplot function with Python.

Additional file 11: Fig. S7. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the first random instantiation.

Additional file 12: Fig. S8. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the second random instantiation.

Additional file 13: Fig. S9. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the third random instantiation.

Additional file 14: Fig. S10. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the fourth random instantiation.

Additional file 15: Fig. S11. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the fifth random instantiation.

Additional file 16: Fig. S12. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the sixth random instantiation.

Additional file 17: Fig. S13. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the seventh random instantiation.

Additional file 18: Fig. S14. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the eighth random instantiation.

Additional file 19: Fig. S15. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the ninth random instantiation.

Additional file 20: Fig. S16. METABRIC gene expression data with 446 random genes selected as the feature set and then undergoing unsupervised hierarchical clustering using Seaborn's clustermap function within Python. This is the tenth random instantiation.

Additional file 21: Fig. S17. PCA plot of Z-scaled data for METABRIC gene expression data, and also MMTV-Myc, MMTV-Neu, and MMTV-PyMT primary mammary mouse tumors and visualized using a matplotlib scatterplot in Python. The first principal component accounts for 44.6% of the variance within the dataset, with three distinct clusters that correspond to different sampling groups appearing. The second principal component accounts for 8.6% of variance within the dataset.

Additional file 22: Fig. S18. PCA plot of Z-scaled and batch corrected METABRIC, MMTV-Myc, MMTV-Neu, and MMTV-PyMT primary mammary tumor gene expression data. Batch correction was performed using ComBat with settings defined in the main text methods section. Principal components are plotted using a scatterplot in matplotlib within Python. The first principal component in the batch corrected dataset accounts for 5.9% of variance, while the second principal component accounts for 4.5% of variance.

Additional file 23: Fig. S19. Recursive feature elimination with 10-fold cross-validation (RFECV) using a support vector machine (SVM) classifier with radial basis function (RBF) kernel was performed on Z-scored and batch corrected METABRIC gene expression data—all mouse data was removed after batch correction and Z-scoring but before RFECV. Gene expression data was limited to the remaining PAM50 genes after combining human and mouse datasets, which was 45 genes. The vertical dashed line was placed at the optimal number of features, which is the number of features which give the highest average accuracy score over the 10 iterations the test is performed. The light blue shaded area represents one standard deviation away from the mean accuracy score. RFECV was performed using the Yellowbrick package in Python.

Additional file 24: Fig. S20. Non-redundant TCGA breast cancer patient data for OBSCN mutation status underwent Kaplan-Meier analysis for overall survival with 95% confidence intervals shown in light blue for the unaltered population and light red for the OBSCN mutated population. KM plots were created using the survminer R package. All breast cancer datasets with mutation data available were analyzed with redundant patients removed from the dataset and the first event occurrence by date for each patient was kept.

Additional file 25: Fig. S21. Volcano plot of differentially expressed genes of MMTV-Myc EMT tumors compared to microacinar and squamous tumors. Log fold change is displayed along the x-axis and -log base 10 of the p-value determined by a students' t-test is displayed along the y-axis. Log fold change cut off set at ≥ 3 and negative log base 10 cut off set at ≥ 1.3, which is equivalent to a p-value ≤ 0.05. Select genes are highlighted. The volcano plot is made using bioinfokit as a Python package.

Additional file 26: Table S1. Estimated integer copy number gain or loss by gene for all 9 tumors that underwent whole genome sequencing.

Additional file 27: Table S2. ssGSEA values for all available EMT, squamous, and microacinar tumors. Includes test statistics, p-values, and log fold changes between groups.

Additional file 28: Table S3. Gene expression and estimated integer copy number correlation values by gene with location included in the mm10 genome across all EMT, squamous, and microacinar samples. Includes the Kendall rank correlation coefficient and Pearson correlation coefficient values.

Additional file 29: Table S4. Integrative mouse gene set found by taking all genes that were differentially expressed between EMT, squamous, and microacinar histological subtypes that also had copy number differences present from the whole genome sequencing data. The rationale behind this is that if the MMTV-Myc histological subtype preferentially represent different human intrinsic breast cancer subtypes, the combination of genomic and transcriptomic data will limit the number of genes used in unsupervised hierarchical clustering to those that matter most.

Additional file 30: Table S5. Contingency table of TCGA breast cancer patient copy number data by select genes available through cBioPortal. An event for each gene is the presence of either a copy number gain or copy number loss. Separate pairwise contingency tables are listed for all possible combinations of genes considered: ERBB2, MYC, COL1A1, KRAS, and FGFR2. Odds ratios and p-values were determined using Fisher's exact test for each gene combination.

Additional file 31: Table S6. Data for constructing the Kaplan-Meier plot of MYC amplified and unaltered breast cancer patients by probability of overall survival in months. Includes bounds for 95% confidence intervals, number at risk, and standard error calculations.

Additional file 32: Table S7. Raw data obtained from cBioPortal for all breast cancer patients separated by MYC amplification status. Redundant patient samples have not been removed.

Additional file 33: Table S8. Data for constructing the Kaplan-Meier plot of KRAS with amplification/activating mutation and unaltered breast cancer patients by probability of overall survival in months. Includes bounds for 95% confidence intervals, number at risk, and standard error calculations.

Additional file 34: Table S9. Raw data obtained from cBioPortal for all breast cancer patients separated by KRAS amplification/mutation status. Redundant patient samples have not been removed.

Additional file 35: Table S10. Data for constructing the Kaplan-Meier plot of FGFR2 amplified and unaltered breast cancer patients by probability of overall survival in months. Includes bounds for 95% confidence intervals, number at risk, and standard error calculations.

Additional file 36: Table S11. Raw data obtained from cBioPortal for all breast cancer patients separated by FGFR2 amplification status. Redundant patient samples have not been removed.

Additional file 37: Table S12. List of the 32 most important genes as a subset of PAM50 for distinguishing between intrinsic subtypes of breast cancer as determined through RFECV using an SVM classifier with RBF kernel.

Additional file 38: Table S13. Data for constructing the Kaplan-Meier plot of OBSCN mutated and unaltered breast cancer patients by probability of overall survival in months. Includes bounds for 95% confidence intervals, number at risk, and standard error calculations.

Additional file 39: Table S14. Raw data obtained from cBioPortal for all breast cancer patients separated by OBSCN mutation status. Redundant patient samples have not been removed.

Additional file 40: Table S15. Sample level enrichment table for genes that are often co-mutated with OBSCN. Samples are ordered from highest to lowest for genes with mutations that occur in the altered group (OBSCN mutated group).

Additional file 41: Table S16. Human gene homologs and genomic location that correspond to genes contained within the conserved microacinar tumor amplification events on mouse chromosomes 11 and 15.

Additional file 42: Table S17. Fisher exact test of ERBB2 and MYC amplification in human breast cancer patients. MYC and ERBB2 amplification trend toward co-occurrence.

Acknowledgements
The authors thank Jesus Alberto Garcia-Lerena, John Vusich, Marcelio Shammami, Reham Ammar, and James Lord of Michigan State University for insightful discussion of various attributes of the manuscript.

Author contributions
CDB wrote the manuscript, performed whole genome DNA extraction and processing of MMTV-Myc samples, performed all bioinformatic analyses, performed all statistical analyses, performed all machine learning, and contributed to the theoretical design of the study. MMOO performed the mutation verification by Sanger sequencing and ssGSEA processing. MSM contributed to the theoretical and technical design of the machine learning components. ERA performed MMTV-Myc tumor extraction and preservation, performed microarrays on mouse transcriptional data, and lead the theoretical design of the study. All authors contributed to the revision of the manuscript.

Funding
Whole genome sequencing of MMTV-Myc tumors was financially possible from an internal MSU and Illumina joint discovery grant. Stipend support for the first author during the course of this study was provided under a US Department of Defense CDMRP #W81XWH-21-1-0002 and a fellowship from the Aitch Foundation, a 501(c)(3) charitable organization located in Lansing, MI. Funding sources had no impact on the study design, sample collection, data analysis, data interpretation, or manuscript preparation.

Availability of data and materials
Mouse MMTV-Myc whole genome sequence data obtained in this study is available under BioProject PRJNA945899. Python and R code used in this study to process data and generate figures is available on GitHub post publication (https://github.com/CarBroke). Other data related to primary and supplementary figure generation are included in supplemental materials.

Declarations

Ethics approval and consent to participate
No human subjects or clinical specimens were involved in this study. All mouse investigations were approved by the institutional animal care and use committee (IACUC) at MSU.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Department of Biochemistry and Molecular Biology, Michigan State University, 567 Wilson Road, BPS Room 2120, East Lansing, MI 48824, USA. [2]Genetics and Genomics Science Program, Michigan State University, 567 Wilson Road, BPS Room 2120, East Lansing, MI 48824, USA. [3]Department of Computational Mathematics, Science, and Engineering, Michigan State University, 428 South Shaw Lane, Engineering Building Room 1508C, East Lansing, MI 48824, USA. [4]Department of Chemical Engineering and Materials Science, Michigan State University, 428 South Shaw Lane, Engineering Building Room 1508C, East Lansing, MI 48824, USA. [5]Department of Physiology, Michigan State University, 567 Wilson Road, BPS Room 2194, East Lansing, MI 48824, USA.

Received: 25 April 2023 Accepted: 30 September 2023
Published online: 07 October 2023

References

1. Xu J, Chen Y, Olopade OI. MYC and breast cancer. Genes Cancer. 2010;1:629–40.
2. Walhout AJM, Gubbels JM, Bernards R, van der Vliet PC, Timmers HTh, M. c-Myc/Max heterodimers bind cooperatively to the E-box sequences located in the first intron of the rat ornithine decarboxylase (ODC) gene. Nucl Acids Res. 1997;25:1493–501.
3. Dang CV. MYC on the path to cancer. Cell. 2012;149:22–35.
4. Perna D, et al. Genome-wide mapping of Myc binding and gene regulation in serum-stimulated fibroblasts. Oncogene. 2012;31:1695–709.
5. Blancato J, Singh B, Liu A, Liao DJ, Dickson RB. Correlation of amplification and overexpression of the c-myc oncogene in high-grade breast cancer: FISH, in situ hybridisation and immunohistochemical analyses. Br J Cancer. 2004;90:1612–9.
6. Deming SL, Nass SJ, Dickson RB, Trock BJ. C-myc amplification in breast cancer: a meta-analysis of its occurrence and prognostic relevance. Br J Cancer. 2000;83:1688–95.
7. Casciano JC, et al. MYC regulates fatty acid metabolism through a multigenic program in claudin-low triple negative breast cancer. Br J Cancer. 2020;122:868–84.
8. Chandriani S, et al. A core MYC gene expression signature is prominent in basal-like breast cancer but only partially overlaps the core serum response. PLoS ONE. 2009;4: e6693.
9. Lin CY, et al. Transcriptional amplification in tumor cells with elevated c-Myc. Cell. 2012;151:56–67.
10. Alluri P, Newman L. Basal-like and triple negative breast cancers: searching for positives among many negatives. Surg Oncol Clin N Am. 2014;23:567–77.
11. Stewart TA, Pattengale PK, Leder P. Spontaneous mammary adenocarcinomas in transgenic mice that carry and express MTV/myc fusion genes. Cell. 1984;38:627–37.

12. D'Cruz CM, et al. c-MYC induces mammary tumorigenesis by means of a preferred pathway involving spontaneous Kras2 mutations. Nat Med. 2001;7:235–9.
13. Andrechek ER, et al. Genetic heterogeneity of Myc-induced mammary tumors reflecting diverse phenotypes including metastatic potential. Proc Natl Acad Sci USA. 2009;106:16387.
14. Sakamoto K, Schmidt JW, Wagner K-U. Mouse Models of Breast Cancer. Methods Mol Biol Clifton NJ. 2015;1267:47–71.
15. Moody SE, et al. Conditional activation of Neu in the mammary epithelium of transgenic mice results in reversible pulmonary metastasis. Cancer Cell. 2002;2:451–61.
16. Gunther EJ, et al. Impact of p53 loss on reversal and recurrence of conditional Wnt-induced tumorigenesis. Genes Dev. 2003;17:488–501.
17. Guy CT, Cardiff RD, Muller WJ. Induction of mammary tumors by expression of polyomavirus middle T oncogene: a transgenic mouse model for metastatic disease. Mol Cell Biol. 1992;12:954–61.
18. Hollern DP, Andrechek ER. A genomic analysis of mouse models of breast cancer reveals molecular features ofmouse models and relationships to human breast cancer. Breast Cancer Res. 2014;16:R59.
19. Pfefferle AD, et al. Transcriptomic classification of genetically engineered mouse models of breast cancer identifies human subtype counterparts. Genome Biol. 2013;14:R125.
20. Manning HC, Buck JR, Cook RS. Mouse models of breast cancer: platforms for discovering precision imaging diagnostics and future cancer medicine. J Nucl Med. 2016;57:60S-68S.
21. Wu S, Zhu W, Thompson P, Hannun YA. Evaluating intrinsic and non-intrinsic cancer risk factors. Nat Commun. 2018;9:3490.
22. Schmidt DR, et al. Metabolomics in cancer research and emerging applications in clinical oncology. CA Cancer J Clin. 2021;71:333–58.
23. Pagès F, et al. Immune infiltration in human tumors: a prognostic factor that should not be ignored. Oncogene. 2010;29:1093–102.
24. Fisher B, et al. Tamoxifen for prevention of breast cancer: Report of the national surgical adjuvant breast and bowel project P-1 study. JNCI J Natl Cancer Inst. 1998;90:1371–88.
25. Piccart-Gebhart MJ, et al. Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. N Engl J Med. 2005;353:1659–72.
26. Weinstein JN, et al. The cancer genome atlas pan-cancer analysis project. Nat Genet. 2013;45:1113–20.
27. Rennhack JP, et al. Integrated analyses of murine breast cancer models reveal critical parallels with human disease. Nat Commun. 2019;10:3261.
28. Ross C, et al. The genomic landscape of metastasis in treatment-naïve breast cancer models. PLoS Genet. 2020;16: e1008743.
29. Campbell KM, et al. A spontaneous aggressive ERα+ mammary tumor model is driven by Kras activation. Cell Rep. 2019;28:1526-1537.e4.
30. Swiatnicki MR, et al. Elevated phosphorylation of EGFR in NSCLC due to mutations in PTPRH. PLOS Genet. 2022;18: e1010362.
31. Hollern DP, et al. E2F1 drives breast cancer metastasis by regulating the target gene FGF13 and altering cell migration. Sci Rep. 2019;9:10718.
32. Andrechek ER. HER2/Neu tumorigenesis and metastasis is regulated by E2F activator transcription factors. Oncogene. 2015;34:217–25.
33. Babraham Bioinformatics - FastQC A Quality Control tool for High Throughput Sequence Data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.
34. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114–20.
35. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. https://doi.org/10.48550/arXiv.1303.3997 (2013).
36. Picard Tools - By Broad Institute. https://broadinstitute.github.io/picard/.
37. Li H, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
38. McKenna A, et al. The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.
39. Koboldt DC, et al. VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 2012;22:568–76.
40. Cingolani P, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms. SnpEff Fly (Austin). 2012;6:80–92.
41. Keane TM, et al. Mouse genomic variation and its effect on phenotypes and gene regulation. Nature. 2011;477:289–94.
42. Rausch T, et al. DELLY: structural variant discovery by integrated paired-end and split-read analysis. Bioinformatics. 2012;28:i333–9.
43. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: a probabilistic framework for structural variant discovery. Genome Biol. 2014;15:R84.
44. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. PLOS Comput Biol. 2016;12: e1004873.
45. Reich M, et al. GenePattern 2.0. Nat Genet. 2006;38:500–1.
46. Krzywinski M, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19:1639–45.
47. Pedregosa F, et al. Scikit-learn: machine learning in python. J Mach Learn Res. 2011;12:2825–30.
48. Rosenthal R, McGranahan N, Herrero J, Taylor BS, Swanton C. deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. Genome Biol. 2016;17:31.
49. Bedre, R. reneshbedre/bioinfokit: Bioinformatics data analysis and visualization toolkit. (2022). 10.5281/zenodo.3964972
50. Li Y, et al. Patterns of somatic structural variation in human cancer genomes. Nature. 2020;578:112–21.
51. Baslan T, et al. Novel insights into breast cancer copy number genetic heterogeneity revealed by single-cell genome sequencing. Life. 2020;9:e51480.
52. Staaf J, et al. High-resolution genomic and expression analyses of copy number alterations in HER2-amplified breast cancer. Breast Cancer Res. 2010;12:R25.
53. Ulz P, et al. Whole-genome plasma sequencing reveals focal amplifications as a driving force in metastatic prostate cancer. Nat Commun. 2016;7:12008.
54. Shao X, et al. Copy number variation is highly correlated with differential gene expression: a pan-cancer study. BMC Med Genet. 2019;20:175.
55. Ohshima K, et al. Integrated analysis of gene expression and copy number identified potential cancer driver genes with amplification-dependent overexpression in 1,454 solid tumors. Sci Rep. 2017;7:641.
56. Rivas MA, et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. Science. 2015;348:666–9.
57. Nik-Zainal S, et al. Mutational processes molding the genomes of 21 breast cancers. Cell. 2012;149:979–93.
58. Alexandrov LB, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578:94–101.
59. Zhan L, et al. Deregulation of Scribble promotes mammary tumorigenesis and reveals a role for cell polarity in carcinoma. Cell. 2008;135:865–78.
60. Paschka P, et al. Adverse prognostic significance of KIT mutations in adult acute myeloid leukemia with inv(16) and t(8;21): A cancer and Leukemia Group B Study. J Clin Oncol. 2006;24:3904–11.
61. di Masi A, et al. Retinoic acid receptors: from molecular mechanisms to cancer therapy. Mol Aspects Med. 2015;41:1–115.
62. Sievers F, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.
63. Kam RKT, Deng Y, Chen Y, Zhao H. Retinoic acid synthesis and functions in early embryonic development. Cell Biosci. 2012;2:11.
64. Wang G, Tian Y, Hu Q, Xiao X, Chen S. PML/RARa blocks the differentiation and promotes the proliferation of acute promyelocytic leukemia through activating MYB expression by transcriptional and epigenetic regulation mechanisms. J Cell Biochem. 2019;120:1210–20.
65. Ornitz DM, Itoh N. The fibroblast growth factor signaling pathway. Wiley Interdiscip Rev Dev Biol. 2015;4:215–66.
66. Curtis C, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.
67. Rueda OM, et al. Dynamics of breast-cancer relapse reveal late-recurring ER-positive genomic subgroups. Nature. 2019;567:399–404.
68. Pereira B, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat Commun. 2016;7:11479.
69. Horiuchi D, et al. MYC pathway activation in triple-negative breast cancer is synthetic lethal with CDK inhibition. J Exp Med. 2012;209:679–96.
70. Bertucci F, Finetti P, Birnbaum D. Basal breast cancer: a complex and deadly molecular subtype. Curr Mol Med. 2012;12:96.
71. Yin L, Duan J-J, Bian X-W, Yu S. Triple-negative breast cancer molecular subtyping and treatment progress. Breast Cancer Res. 2020;22:61.

72. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8:118–27.
73. Parker JS, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–7.
74. van der Maatern L, Hinton G. Visualizing data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
75. Fougner C, Bergholtz H, Norum JH, Sørlie T. Re-definition of claudin-low as a breast cancer phenotype. Nat Commun. 2020;11:1787.
76. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. BMC Genom. 2020;21:6.
77. Newton EE, Mueller LE, Treadwell SM, Morris CA, Machado HL. Molecular targets of triple-negative breast cancer: Where do we stand? Cancers. 2022;14:482.
78. Skoulidis F, et al. Sotorasib for Lung Cancers with KRAS p.G12C Mutation. N Engl J Med. 2021;384:2371–81.
79. Jänne PA, et al. Adagrasib in non–small-cell lung cancer harboring a KRASG12C mutation. N Engl J Med. 2022;387:120–31.
80. Coombes RC, et al. Results of the phase IIa RADICAL trial of the FGFR inhibitor AZD4547 in endocrine resistant breast cancer. Nat Commun. 2022;13:3246.
81. Campbell BB, et al. Comprehensive analysis of hypermutation in human cancer. Cell. 2017;171:1042-1056.e10.
82. Creighton CJ. The molecular profile of luminal B breast cancer. Biol Targets Ther. 2012;6:289–97.
83. Prat A, et al. Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res. 2010;12:R68.
84. Shadeo A, Lam WL. Comprehensive copy number profiles of breast cancer cell model genomes. Breast Cancer Res. 2006;8:R9.
85. Pariyar M, Johns A, Thorne RF, Scott RJ, Avery-Kiejda KA. Copy number variation in triple negative breast cancer samples associated with lymph node metastasis. Neoplasia. 2021;23:743–53.

Publisher's Note