Methods for

Multiplatform Analysis of Primary and Metastatic Breast Tumors from the AURORA US Network identifies microenvironment and epigenetics differences as drivers of metastasis

Clinical Summary

Samples from a total of 55 patients with metastatic breast cancer were the final data set of the AURORA US cohort. Of these 55 women, 10 (18%) were of African American descent and 4 (7%) were of Hispanic ethnicity. Median age at initial breast cancer diagnosis was 49 years (range: 25-76). Forty-nine patients (89%) initially presented with stage I-III breast cancer, of which 19 (38%) received neoadjuvant systemic therapy, and six patients (10%) presented with de novo metastatic disease. Ductal histology was most prevalent among the cohort (n=44, 80%); 7 patients (12%) were diagnosed with lobular or mixed lobular/ductal carcinoma. The distribution of breast cancer receptor subtype per clinical testing at initial diagnosis was triple-negative, n=19 (34%); hormone receptor (HR)-positive/HER2-negative, n=17 (30%); HR-positive/HER2-positive, n=6 (10%); HR-negative/HER2-positive, n=4 (7%); and unknown, n=9 (16%). In the metastatic setting, patients received a median of 3 lines of systemic therapy (range: 0-20). Metastatic samples from a total of 20 patients were collected at autopsy. Additional clinicopathologic features are displayed in Supplementary table 1.

Pathology Review

Pathology quality control was performed on each tumor specimen and normal tissue specimen as an initial QC step. Hematoxylin and Eosin (H&E) stained sections from each sample were subjected to independent pathology review to confirm that the tumor specimen was histologically consistent to the reported histology. The percent rumor nuclei, percent necrosis, and other pathology annotations were also assessed. Tumor samples with ≥30% tumor nuclei, and normal tissue with 0% tumor nuclei, were submitted for nucleic acid extraction. All H&E images are also available and part of this data resource.

AURORA Sample acquisition and Biospecimen Processing

RNA and DNA were extracted from frozen tissues using a modification of the AllPrep DNA/RNA kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a mirVana miRNA Isolation Kit (Ambion). RNA and DNA were extracted from FFPE solid tissues using a modification of the AllPrep DNA/RNA FFPE kit (Qiagen). The flow-through from the Qiagen DNA column was processed using a mirVana miRNA Isolation Kit (Ambion). For cases in which whole blood or blood derivatives were received, DNA was extracted from blood using the QiaAmp DNA Blood Midi kit (Qiagen). RNA samples were quantified by measuring Abs260 with a UV spectrophotometer and DNA quantified by PicoGreen assay. DNA specimens were resolved by 1% agarose gel electrophoresis to confirm high molecular weight fragments. A custom Sequenom SNP panel or the AmpFISTR Identifiler (Applied Biosystems) was utilized to verify that tumor DNA and germline DNA representing a case were derived from the same patient. RNA was analyzed via the RNA6000 Nano assay (Agilent) for determination of an RNA Integrity Number (RIN). Only cases yielding a minimum of 500ng of tumor DNA, 500ng of tumor RNA, and 500ng of germline DNA were included in this study. A minimum of one QC qualified tumor sample and a QC qualified normal were required for a case to become part of the study (n=55 total cases).

RNA sequencing, gene expression data values and normalization

Gene expression profiles from primary and metastatic tumors for AURORA dataset were generated by RNA-sequencing using an Illumina HiSeq and a rRNA-depletion method. Briefly, 300-500ng total RNA was converted to RNAseq libraries using the TruSeq Stranded Total RNA Library Prep Kit with Ribo-Zero Gold (Illumina) and sequenced on an Illumina HiSeq 2000 using a 2x50bp configuration. Quality-control-passed reads were aligned to the human reference CGRh38/hg38 genome using STAR[1]. Transcript abundance estimates for each sample were performed using Salmon[2], an expectation-maximization algorithm using the UCSC gene definitions. Raw read counts for all RNAseq samples were normalized to a fixed upper quartile (UQN)[3]. The raw reads files are available in dbGAP (submission in process)

Gene expression analysis of RNAseq data and batch effect adjustments

RNAseq UQN gene counts from 123 primary and metastatic tumors comprised of 35 FFPE and 88 Fresh Frozen (FF) RNA-sequenced tumor data were log2 transformed, genes were filtered for those expressed in 70% of samples and zeros were returned to the empty values. In order to improve the batch effect between the two data types (i.e. FFPE vs FF), we merged a second dataset of 101 paired primary and metastatic tumors (UNC Rapid Autopsy donation Program (RAP) cohort) comprised of 20 FFPE and 81 FF sequenced tumors. This second dataset was partially previously published in 2018[4], but some new samples were added and sequenced for the present work, and many of the published samples were resequenced here using the rRNA-depletion method (dbGAP phs002429). The RAP 101 samples of the present work were created with the same RNA extraction, library preparation and sequencing protocol as are AURORA samples, and represents a second data set of FFPE and FF samples that increases our sample size for adjustments of FFPE vs FF effects; note that the RAP101 set is also a second data set of primary tumor and metastasis pairs as well. The clinical information of the RAP101 dataset is found in Supplementary table 2.

To address this systematic effect, we merged the raw read counts for all RNAseq samples of the previously mentioned RAP 101 dataset with 123 samples of AUORA study (Level 1 data). These counts were normalized using DESeq2-normalized counts (median of ratios method)[5]. Briefly, we created DESeq2Dataset object and generated size factors using estimateSizeFactors() function. Next, to retrieve the normalized counts matrix, we used the counts () function and add the argument normalized=TRUE. After generating the normalized count matrix, genes with an average expression less than 10 were filtered from the dataset. RNAseq normalized gene counts from the 224 dataset was log2 transformed (Level 2 data). Next, we used the removeBatchEffect () function from limma R package[6] including both batches in the formula. Lastly, we subtracted only the 123 samples from the AURORA study and used this normalized, log2 transformed and batch corrected dataset for further RNAseq gene expression analysis (Level 3 data).

In order to minimize false positive results due to the normal tissue contamination generated by normal brain (n=10), liver (n=8) or lung tissue (n=7), the most common sites of metastasis in this study, we removed those genes whose expression was solely coming from these three tissue sites. Specifically, we used supervised leaning to determine a normal brain, liver and lung signature from comparing each normal tissue vs normal breast tissue (n=5) (Supplementary table

3, dbGAP accession number (submission in process) for AURORA and (phs002429) for RAP and 9830). This normal tissue dataset was also created using the same RNA extraction, library preparation and sequencing protocols. From normalized, filtered and median counts we performed linear model (LM) regression using lme4[7] and lmerTest[8] R package given the formula, Fit = lm(Genes~Normal site of site of metastasis/Breast normal) and p-values were adjusted for multiple comparisons using the Holm-Bonferroni. We obtained the most significant upregulated genes each normal tissue (FDR< 0.00001) from comparing each normal tissue vs normal breast tissue (Brain vs. Breast, Liver vs. Breast and Lung vs. Breast); we took and merged these 3 lists and identified 1900 genes as the distinctive upregulated genes of our "normal tissue signature". In order to build a second signature characteristic of breast primary tumors, we did a second LM analysis between the 46 primary tumors from AURORA study and the 5 normal breast tissue from the above-mentioned normal tissue cohort and we obtained 833 significant upregulated genes (FDR<0.01). Some of these genes were also present in the "normal tissue signature" and thus we removed these common 449 genes from the "normal tissue signature" list considering these genes not unique for normal tissues but also being important markers for primary tumors in the AURORA cohort. Finally, the remaining 1451 genes of the "normal tissue signature" (Supplementary table 3) were removed from the original normalized and batch corrected gene expression data matrix of the 123 AURORA cohort (referred to hereafter as the normalized, log2 transformed and batch corrected and normal-adjusted data, or Level 4 RNAseq data).

PAM50 subtype classification

In order to better maintain methods with past intrinsic subtyping methods[9-11], for PAM50 subtype classification assignments we normalized the RNAseq data in a different way than described immediately above, and that is based upon within data set row and column standardizations. Briefly, RNAseq normalized gene counts from 123 primary and metastatic tumors comprised of 35 FFPE and 88 Fresh Frozen (FF) RNA-sequenced tumor data were log2 transformed, genes were filtered for those expressed in 70% of samples and zeros were returned to the empty values. To address the FFPE vs FF effects, we again used the AURORA and RAP101 data sets as described above and made an adjustment for FFPE vs FF. Namely, using only common genes between both datasets, we merged, row median centered and column standardized separately FFPE and FF groups, where each gene was a row, and each sample was a column. Next, we subtracted only the FFPE and FF normalized batches from AURORA study only and used these values for ROC curve and Youden cut-off analysis for ER, PR, and HER2 status comparisons, which provide external validation that the adjustments do not adversely affect the gene expression data using tests of correlation to the external clinical standards. For PAM50 subtype classification we applied a HER2/ER subgroup-specific gene centering method as described in the Supplemental Methods of Fernandez-Martinez et al.[10]. For applying this subgroup-specific gene-centering method, we need the IHC status for all samples assayed by RNAseq. 6% of primary tumors and 39% of metastatic samples did not have HER2 IHC information, and 38% of metastatic samples were missing for ER status. “Profiled Primary ER/HER2/PR columns of Supplementary table 2 were used for this analysis. We again used ROC curve and Youden cut-off value for inferring protein clinical status using ESR1 and ERBB2 gene expression data from all tumors, and we assigned ER and HER2 clinical status to those samples that had missing clinical values using the mRNA surrogates. The ROC curve analysis showed 0.92 value for ER status by ESR1 mRNA, and 0.86 for HER2 status using ERBB2 mRNA. These new RNAseq inferred ER/PR/HER2 protein status were used for the subgroup-specific gene centering method (“Inferred ER/PR/HER2 column of Supplementary table 2). Finally, the gene expression values of the PAM50 genes using the UQN gene counts were then normalized and then the PAM50 predictor[12] was applied using the provide centroids, to assign subtype calls using correlation values for all primary tumors and metastases (Supplementary table 2).

Gene expression signatures

For each batch corrected and adjusted for normal tissue gene expression data set/subset (Level 4 RNAseq data), we applied a collection of 747 gene expression modules (Supplementary table 3), representing multiple biological pathways and cell types, to all primary and metastatic tumors. 701 signatures were obtained from 125 publications partially summarized previously [13-15] and 48 Gene set enrichment analysis (GSEA) signatures published in the Molecular Signature Database[16]. In detail, 1) 669 modules were calculated as the median value of each gene expression value present in the signature for each sample of the set used; 2) 20 were the value of a single gene; and 3) 57 named as “special modules” that used specific predetermined algorithms previously described [9,17-37] (in order to implement each modules, the methods detailed in the original studies were followed as closely as possible).

Finally, we newly developed an immune metagene signature named “GP2 Immune-Metagene”, signature which we developed to capture immune cells features as derived from the AURORA data set. Briefly, we used TCGA gene expression data to calculate all our 747 module scores, which was then used for hierarchical cluster analysis, and the resulting clusters of modules then tested for significance of these groups of modules using SigClust[38]. 56 Clusters with a $p<0.001$ were identified and 16 immune related signatures from cluster 51 were grouped as a new “immune meta-signature” named GP2 Immune-Metagene signature (Supplementary table BB); included within this group of immune clusters were signatures of Tcells, Bcells, Macrophages, and Dendritic cells. Next, using our previously calculated 747 gene expression modules scores from AURORA dataset we selected the 16 immune related signatures and calculated the mean of these 16 signatures for each patient and called this new derived signature as “GP2 Immune Metagene”.

Merging UNC Rapid Autopsy donation Program (RAP), GEICAM/2009-03 ConvertHER trial (GEICAM) and AURORA cohorts

To create as large a data set as possible, we merged the data of the AURORA, RAP101, and 204 samples of GEICAM/2009-03 ConvertHER trial (GEICAM cohort)[11]; this yielded a final cohort of 428 tumors in total (158 patients with 159 primaries and 400 paired metastasis, 17 unpaired primaries and 11 unpaired metastasis), summarized in Supplementary table 2. RNAseq-Sequencing data of 204 GEICAM study were retrieved from dbGaP, accession number phs001866, and the processed data in GEO (GSE147322).

Next, we corrected the technical bias detected between the gene expression of 259 FFPE and 169 Fresh frozen (FF) samples from 176 primary and 411 metastatic tumors. The raw counts of the 428 tumors were normalized using DESeq2-normalized counts (median of ratios method)[5]. We created DESeq2Dataset object and generated size factors using estimateSizeFactors() function. Next, to retrieve the normalized counts matrix, we used the counts() function and add the argument normalized=TRUE. After generating the normalized count matrix, genes with an average expression less than 10 were filtered from the dataset. RNAseq normalized gene counts from the 428 tumors were log2 transformed. Next, we used the removeBatchEffect () function from limma R package[6] indicating FFPE or FF as batches in the formula (removeBatchEffect (normlog2data, batch). In order to minimize the false positive results due to the normal tissue contamination we proceed as we did in AURORA dataset, 1451 genes of the "normal tissue signature" (Supplementary table 3) were removed from the data matrix of the 428 AURORA-RAP-GEICAM cohort to minimize the false positive results coming from normal tissue contamination.

Next on the 3 data set combined data matrix, we calculated the gene signature score for each module as described before, and we performed linear mixed model (LMM) using lmerTest and lme4 R package to identify significantly changed modules between metastatic and primary tumors. In the linear model we included the term "patient" as random effect or cofounding variable: Fit = lmer(Genes~ Met/Prim + (1|Patient) using all the primary and metastatic tumors except the primaries identified as post-treatment primaries (patients who received neo-adjuvant therapy prior to primary tumor collection). To avoid the possible confounding factor of intrinsic molecular subtype in the subsequent analysis, we divided tumors into two datasets based upon the subtype of the primary tumor from each pair: a "luminal set" comprising all Luminal A, Luminal B and HER2E subtype patients and a "basal-like set" containing basal-like subtype only; samples called normal-like in either the primary or metastatic tumors or post-treatment primary tumors were removed from the analysis (column "Groups PAM50 Gene expression analysis" from supplementary table 2). To identify significantly changed modules between brain or liver and their corresponding primary tumors only the studied sites of metastasis versus the corresponding primary pair were compared using the same lmer function. The significant differentially expressed modules (FDR<0.05) were hierarchically clustered using ComplexHeatmap R package. HeatmapAnnotation and Heatmap functions were used to show the heatmap that was previously row ordered by primary and metastatic tumors and column ordered by estimates or beta values. Differential gene expression modules analysis in the merged AURORA-RAP-GEICAM set were performed in the same way than AURORA only. Multi-metastatic samples derived from AURORA and RAP and single primary-tumor pairs derived from GEICAM with PAM50 classification of Normal-like in primary or metastatic tumors and post-treatment primary tumors were removed from the analysis. For the comparisons between site of metastasis using the merged set, we performed SAM[39] analysis and the differentially expressed modules (FDR=0) between 48 Liver metastasis vs 21 Brain metastasis, 48 Liver metastasis vs 27 Lung metastasis, 48 Liver metastasis vs 27 Lung metastasis, 48 Liver metastasis vs 38 LN metastasis, 21 Brain metastasis vs 38 LN metastasis and 27 Lung metastasis vs 21 Brain metastasis (Supplementary table 3).

Statistical Methods

For Linear Mixed Models/ Linear Mixed Effects Model and Linear Models analysis between primary and metastatic tumors the lmerTest[8] package summary includes coefficient table with estimates and p-values for t-statistics using Satterthwaite's method. These p-values were adjusted for multiple comparisons using the Holm-Bonferroni approach[40]. Nonparametric, two-sided, exact tests were used to make comparisons. A Mann-Whitney U test was used for comparisons between different groups, and a Paired t-test was used for analyzing repeated measures within the same groups. Correlations were measured using the Pearson or Spearman correlation coefficient.

TCGA RNAseq data

We analyzed the breast cancer dataset from The Cancer Genome Atlas (TCGA) project profiled using the Illumina HiSeq system. We included 1095 primary tumors and 97 adjacent non-malignant tissues for developing the immune signature named "GP2 Immune-Metagene" and 761 primary tumors and 74 adjacent non-malignant tissues for the HLA-A methylated primary tumors analysis and prognostic value of HLA-A. TCGA files were downloaded from Broad GDAC Firehose: (https://gdac.broadinstitute.org/runs/stddata__latest/data/BRCA/20160128/ "gdac.broadinstitute.org_BRCA.Merge_rnaseq__illuminahiseq_rnaseq__unc_edu__Level_3__gene_expression__data.Level_3.2016012800.0.0.tar.gz").

Array-based DNA methylation assay

DNA methylation was evaluated using the Illumina HumanMethylationEPIC (EPIC) array (Illumina, CA, USA). The EPIC platform analyzes the DNA methylation status of up to 863,904 CpG loci and 2,932 non-CpG cytosines, spanning gene-associated CpGs as well as a large number of enhancer/regulatory CpGs in intergenic regions[41]. Briefly, DNA was quantified by Qubit fluorimetry (Life Technologies) and 500ng of DNA from each sample was bisulfite-converted using the Zymo EZ DNA Methylation Kit (Zymo Research, Irvine, CA USA) following the manufacturer's protocol using the specified modifications for the Illumina Infinium Methylation Assay. After conversion, all bisulfite reactions were cleaned using the Zymo-Spin binding columns, and eluted in Tris buffer. Following elution, BS converted DNA was processed through the EPIC array protocol. For FFPE samples, the entire BS converted eluate was used as input for the Infinium HD FFPE DNA Restore kit, and processed through the separate restoration workflow. To perform the assay, converted DNA was denatured with NaOH, amplified, and hybridized to the EPIC bead chip. An extension reaction was performed using fluorophore-labeled nucleotides per the manufacturer's protocol.

DNA methylation data packages

DNA methylation data were packaged into four levels as follows.

LEVEL 1: Level 1 data contain raw IDAT files (two per sample with the extensions _Grn.idat and _Red.idat for the two color channels) as produced by the Illumina iScan system. The mapping between IDAT file names and AURORA sample barcodes is provided in Sample.mapping.tsv.

LEVEL 2: Level 2 data contain the signal intensities corresponding to methylated (M) and unmethylated (U) alleles and detection P-values for each probe as extracted by the *readIDATpair* function in the R package *SeSAMe* (https://github.com/zwdzwd/sesame) from the IDAT files. The P-values are calculated using *pOOBAH* (P-value with Out-Of-Band probes for Array Hybridization), which is based on empirical cumulative distribution function of the out-of-band signal from all Type-I probes[42].

LEVEL 3: Level 3 data contain β values defined as $S_M /(S_M+S_U)$ for each locus calculated using the R package *SeSAMe*, where S_M and S_U represent signal intensities for methylated and unmethylated allele. The raw signal intensities are first processed with background correction and dye-bias correction. The background correction is based on the *noob* method[43]. The dye-bias is corrected using non-linear quantile interpolation-based method using the *dyeBiasCorrTypeINorm* function[42]. β values are then computed using the *getBetas* function. Probes with a detection P-value greater than 0.05 in a given sample are masked as NA. Whether the probe is masked due to detection failure is recorded in an extra column (Masked_by_Detection_P_value) to distinguish from experiment-independent masking of probes

(N=105,454) subject to cross-hybridization and genetic polymorphism. The experiment-independent masking is based on the MASK_general column of the file named EPIC.hg38.manifest.tsv (release 20180909) downloaded from http://zwdzwd.github.io/InfiniumAnnotation[41]. From the same source, an additional file (EPIC.hg38.manifest.gencode.v22.tsv) is also included to provide detailed annotation of transcription association for each probe.

LEVEL 4: Level 4 data contain merged data matrix with β values across all samples. Probes masked as NA concerning the probe design in Level 3 data are removed. Sixteen FFPE samples that initially yielded low-quality data were rerun. The resulting two data sets values were merged probe-wise by taking the mean β value. If data was masked in one of the runs, we took available data from the other run.

Nomenclature for control samples:

We include several cell line control samples in each batch to allow for the evaluation of potential batch effects and to facilitate correction of observed batch effects.

Control sample IDs that start with "VARI-Control-" can be interpreted as follows:

VARI-Control-[Batch number]-[Cell line name)-(DNA Isolate ID (A,B,..)]-[Assay Technical Replicate (1,2,3...sequential across batches for the same DNA Isolate)].

External DNA methylation data sets

We processed additional normal tissue DNA methylation data from ENCODE[44] and GEO[45]. We collected raw IDAT files for 24 samples from seven tissue types, including adrenal gland (n=5), liver (n=1), lung (n=4), ovary (n=2), skin (n=4), blood (n=6), and brain (n=2), that were frequently represented as a site of metastasis. We generated β values using the R package *SeSAMe* as described above for the AURORA samples. Further information on these data sets is provided in Supplementary table 4.

Global DNA hypermethylation analysis

To examine cancer-associated DNA hypermethylation profiles, we first used DNA methylation data from normal tissues to eliminate CpG sites that involved in tissue-specific methylation (mean β value > 0.2 in any of the eight tissue types). We eliminated additional CpGs that were significantly differentially methylated between FF and FFPE samples (t-test FDR-adjusted P-value < 0.01 and absolute mean β-value difference > 0.25). For the heatmap analysis shown in Fig.1c, we used 5,000 most variably methylated CpGs across tumors. The probes lacked methylation in the normal tissues (N=146,385) and the subset (N=5,000) used in the heatmap are listed in Supplementary table 4.

Distal element DNA hypomethylation associated with metastasis

We identified 152,211 CpGs in dELSs (distal enhancer-like signatures fall more than 2 kb from the nearest TSS) defined by the ENCODE project[46]. We then selected 19,607 CpGs that are constitutively methylated across eight normal tissue types (mean β value > 0.8). Using the 19,607 CpGs sites, we fitted a probe-wise linear mixed-effects model with terms including primary vs. metastasis, tumor purity, and patient (coded as a random effect) as implemented in the R package *lme4*[47]. P-values were estimated based on the Satterthwaite's approximation method included in the *lmerTest* package in R[47], and adjusted for multiple testing using the Benjamini–Hochberg approach[48]. To examine transcription factors that bind to the CpG sites hypomethylated in metastatic tumors, we analyzed 11,348 ChIP-seq data on 1,359 individual DNA binding factors curated in the Cistrome Data Browser (DB)[49]. The statistical significance of enrichment for transcription factor binding sites among the hypomethylated CpGs was determined using Fisher's exact test with 200bp regions centered on the target CpGs using the R package *LOLA*[50]. All CpGs on the array overlapping the dELSs were used as the background set. P-values were adjusted for multiple comparisons using the Benjamini-Hochberg method.

Putative *ESR1* and *FOXA1* Enhancer Target Genes Affected by Metastasis-Associated DNA Hypomethylation

We identified 47 significantly hypomethylated CpGs overlapping the binding sites for *ESR1* or *FOXA1*.To investigate putative target genes affected by DNA hypomethylation, we first collected 4,681 putative targets of either *ESR1* or *FOXA1* in breast cancers as predicted by Cistrome Cancer[51]. We then considered at most ten nearest genes within 1,000kb upstream and ten nearest genes within 1,000kb downstream from the affected CpG sites, resulting in a list of

121 potential target genes. Gene Ontology GO terms over-representation analysis was performed using the *enrichGO* function with default parameters as implemented in the R package *clusterProfiler* [52].

Identification of DNA hypermethylation associated with metastasis

To identify CpG sites hypermethylated in metastatic tumors compared to primary tumors, we used the 146,385 probes unmethylated in normal tissues defined above. We fitted a probe-wise linear mixed-effects model with terms including primary vs. metastasis, tumor purity, and patient (coded as a random effect) as implemented in the R package *lme4*[7]. P-values were estimated based on the Satterthwaite's approximation method included in the *lmerTest* package[8] in R and adjusted for multiple testing using the Benjamini–Hochberg approach[40].

CpG target analysis

Probes located in the PcG target sites (Fig.5e, j, and o) were determined using H3K27me3 ChIP-seq peaks on the H1 embryonic stem cells generated by the NIH Roadmap Epigenomics Consortium[53]. The broad peaks were downloaded using the R package AnnotaitonHub (ID: AH28888).

TCGA DNA methylation data

We analyzed the breast cancer dataset from The Cancer Genome Atlas (TCGA) project, including 761 primary tumors and 74 adjacent non-malignant tissues profiled using the Infinium HumanMethylation450 (HM450) array. IDAT files were downloaded from the NCI Genomic Data Commons (GDC) Legacy Archive (https://portal.gdc.cancer.gov/legacy-archive)[54], and processed using openSeSAMe pipeline implemented in the R package SeSAMe[42].

DNA sequencing of tumor and normals

Due to variable DNA quality, ranging from high (>2 kb; 131 samples) to medium (0.5-2 kb; 18 samples) and low (<0.5 kb; 44 samples), the 193 AURORA samples were binned into three different batches. For each batch, library construction used the NEBNext UltraII FS DNA Library Prep kit (New England Biolabs, Ipswich, MA) with a 15-minute enzymatic fragmentation. Each library received a unique dual-indexed adapter (Integrated DNA

Technologies, Coralville IA) allowing for both low pass whole genome sequencing (WGS) and multiplex hybrid capture enrichment. Libraries were pooled at 2-4µL, based on final library quality and yield. To evaluate library representation dues to variable DNA quality, we performed survey WGS sequencing for proper library balancing. The pooled libraries were concentrated and diluted to 2.25nM for survey sequencing on the NovaSeq 6000.

Exome hybrid capture utilized the IDT xGen Exome Research Panel v1.0 enhanced with the xGenCNV Backbone Panel-Tech Access (Integrated DNA Technologies, Coralville, IA). The remaining pooled libraries were hybridized to this probe set according to the manufacturer's protocol. The captured products were eluted following precipitation with streptavidin-labeled magnetic beads, amplified by PCR and quantitated prior to dilution and preparatory flow cell amplification for Illumina sequencing. Illumina paired-end sequencing (recipe: 151x17x8x151) performed on the NovaSeq 6000 using the S4 flow cell configuration. For WGS, we targeted 5X coverage, and for WES we aimed for an average unique, on-target sequencing coverage depth of 500X for the tumor and 250X for the matched normal tissue.

Churchill Secondary Analysis for DNA sequencing

The NCH-developed *Churchill* secondary-analysis pipeline[55] was used to process paired-end read data for all samples, utilizing attached UMIs. Reads were aligned to reference genome GRCh38.d1.vd1 via *bwa-mem*, with the resulting alignment deduplicated using GATK's (Picard) *MarkDuplicates* and base scores recalibrated using GATK's *BaseRecalibrator* and *ApplyBQSR*. Variant-calling was then performed on the final deduplicated, recalibrated BAMs. Germline variants were called using GATK's *HaplotypeCaller*; somatic variants were called using GATK's *Mutect2*, with the paired normal samples used to exclude germline variants, and somatic variant filters from *Mutect2* were applied. Additionally, somatic variants from FFPE sources were using corrected variant allele frequency, read start diversity, and unique read ends as indicators of preservation-sourced artifacts. Descriptions of the specific filters can be found below. All SNVs and INDELs were annotated via *SnpEff*, using the GDC.h38 GENCODE v22 database[56]. To ensure samples were of usable quality, depth and breadth metrics were generated by *mosdepth*[57], oxidation and insert size metrics were generated by GATK's

Collect*OxoGMetrics* and *CollectMultipleMetrics* tools, and sequence-usability (duplicate, softclipping, mapq0, unaligned) metrics were generated via *samtools*[58] and custom scripts.

FFPE Filtering

FFPE_filter_LMR_VAF_0.04

Local Mismatch Rate Corrected Variant Allele Frequency below 4%. The local mismatch rate of a variant is the number of mismatched bases in all reads aligned within a 10 bp window each side of the position divided by the total number of bases aligned in this region. This value (LMR) is subtracted from the VAF and if the result is below 4% the variant will be filtered.

FFPE_filter_RSD

Read start diversity filter. The number of unique start positions of all variant supporting reads are counted (after soft trimming). For variants with over 15 supporting reads, at least 4 unique starting positions are required to pass this filter. For variants with over 5 supporting reads, at least 2 unique starting positions are required.

FFPE_filter_URE

Unique Nearest Read End filter. For all variant supporting reads, either the start position or the end position, whichever is closest to the variant (after soft trimming) is recorded. For variants with over 15 supporting reads, at least 4 unique positions are required to pass this filter. For variants with over 5 supporting reads, at least 2 unique positions are required.

CNV/LOH

Copy-number changes and loss-of-heterozygosity events in WGS samples were detected using GATK's *GermlineCNVCaller*[59], with the Churchill pipeline's final BAM alignments as input. Intervals of 1000 bp were used to bin only SNVs found in gnomAD at a frequency of 0.01% or greater. Germline CNV events were identified by comparing individual normal samples to a panel-of-normals composed of all other germline normal samples. Somatic CNV events were identified by comparing each somatic sample for a case to that case's paired germline normal. Following this, CNV events were annotated with the symbols of genes they affected, producing gene-specific copy-ratios.

Additionally, copy number derived from the raw denoised copy ratio signal were produced and plotted across the HLA locus chr6:28,510,120-33,480,577. A smoothing factor was applied by reducing the number of regions into bins by 50-fold and calculating the mean log2 value for each bin. HLA-A/B/C/DRB5 genes were specifically noted for overlap with prominent deletions in the region.

Clonality and Tumor Purity

Clonal variation within and among tumor samples was assessed using *superFreq*[60]. Output BAM alignments from the Churchill pipeline were filtered down to only unique reads overlapping a probe-targeted region. The filtered alignments were then re-genotyped, using *Varscan2*[61] to identify the presence or absence of each of a case's variants in each of its samples. With these inputs, *superFreq* assesses likely copy-number and loss-of-heterozygosity events in combination with SNV and indels to generate the most likely substructure of clones for each sample. The percent composition of tumor cells of all clones was totaled to determine the cellularity of each sample. For each clone, variants in ClinVar- and COSMIC-listed genes are highlighted, as well as mutations of likely-damaging types (frameshift and nonsense); these variants were then queried in the VarSome database, with 'Pathogenic' and 'Likely Pathogenic' variants being considered as potentially consequential clonal variation. Finally, to assess the relationship between clonal diversification patterns and medically-relevant disease characteristics, population genetics and ecological diversity metrics (Fst[62] and Shannon's H[63], respectively) were calculated from clone data via custom scripts.

Neoantigen Prediction

Somatic variants from samples where both DNA and RNA sequencing data were available were evaluated as potential neotantigens using pVACseq, part of the pVACtools package[64]. SNVs and INDELs, after Mutect2 and FFPE filtering when appropriate, were combined with gene expression data to identify and prioritize tumor-specific neoepitopes that are both expressed and has a significantly increased binding affinity compared to the wild-type epitope in the context of the subject's HLA class I alleles. pVACseq's recommended settings and parameters were used for all neoantigen predictions within this cohort.

Resources Table

Resource / Deposited data	**Source**	**Identifier**
AURORA	dbGAP	Submission in progress
TCGA-BRCA mRNA-seq data	Broad GDAC Firehose; dbGAP	https://gdac.broadinstitute.org/runs/stddata__latest/data/BRCA/20160128/; dbGaP accession phs000178
TCGA-BRCA DNA methylation data	NCI GDC	https://portal.gdc.cancer.gov/legacy-archive
UNC Tumor donation program (RAP and 9830)	dbGAP	phs002429
GEICAM/2009-03 ConvertHER trial (GEICAM cohort)	dbGAP; GEO	phs001866; GSE147322

References for Supplemental Methods

1 Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15-21, doi:10.1093/bioinformatics/bts635 (2012).

2 Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat Methods* **14**, 417-419, doi:10.1038/nmeth.4197 (2017).

3 Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. *BMC bioinformatics* **11**, 94 (2010).

4 Siegel, M. B. *et al.* Integrated RNA and DNA sequencing reveals early drivers of metastatic breast cancer. *J Clin Invest* **128**, 1371-1383, doi:10.1172/JCI96153 (2018).

5 Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology* **15**, 550, doi:10.1186/s13059-014-0550-8 (2014).

6 Ritchie, G. R. S., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat Methods.* **11**, doi:10.1038/nmeth.2832 (2014).

7 Douglas Bates, M. M., Ben Bolker, Steve Walker. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, doi:10.18637/jss.v067.i01 (2015).

8 Alexandra Kuznetsova, P. B. B., Rune H. B. Christensen. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* **82**, doi:10.18637/jss.v082.i13 (2017).

9 Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506-519, doi:10.1016/j.cell.2015.09.033 (2015).

10 Fernandez-Martinez, A. *et al.* Survival, Pathologic Response, and Genomics in CALGB 40601 (Alliance), a Neoadjuvant Phase III Trial of Paclitaxel-Trastuzumab With or Without Lapatinib in HER2-Positive Breast Cancer. *J Clin Oncol*, JCO2001276, doi:10.1200/JCO.20.01276 (2020).

11 Garcia-Recio, S. *et al.* FGFR4 regulates tumor subtype differentiation in luminal breast cancer and metastatic disease. *J Clin Invest*, doi:10.1172/JCI130323 (2020).

12 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* **27**, 1160-1167, doi:10.1200/JCO.2008.18.1370 (2009).

13 Fan, C. *et al.* Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. *BMC Med Genomics* **4**, doi:10.1186/1755-8794-4-3 (2011).

14 Gatza, M. L., Silva, G. O., Parker, J. S., Fan, C. & Perou, C. M. An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat Genet* **46**, 1051-1059, doi:10.1038/ng.3073 (2014).

15 Garcia-Recio, S. *et al.* FGFR4 regulates tumor subtype differentiation in luminal breast cancer and metastatic disease. *The Journal of Clinical Investigation* **130**, 4871-4887, doi:10.1172/JCI130323 (2020).

16 Liberzon, A. *et al.* The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst* **1**, 417-425, doi:10.1016/j.cels.2015.12.004 (2015).

17 Broz, M. L. *et al.* Dissecting the Tumor Myeloid Compartment Reveals Rare Activating Antigen-Presenting Cells Critical for T Cell Immunity. *Cancer Cell* **26**, 938, doi:10.1016/j.ccell.2014.11.010 (2014).

18 Prat, A. *et al.* A PAM50-Based Chemoendocrine Score for Hormone Receptor-Positive Breast Cancer with an Intermediate Risk of Relapse. *Clin Cancer Res* **23**, 3035-3044, doi:10.1158/1078-0432.ccr-16-2092 (2017).

19 Prat, A. *et al.* Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. *Breast Cancer Res* **12**, doi:10.1186/bcr2635 (2010).

20 Paik, S. *et al.* A Multigene Assay to Predict Recurrence of Tamoxifen-Treated, Node-Negative Breast Cancer. *New England Journal of Medicine* **351**, 2817-2826, doi:10.1056/NEJMoa041588 (2004).

21 D'Arcy, M. *et al.* Race-associated biological differences among Luminal A breast tumors. *Breast cancer research and treatment* **152**, 437-448, doi:10.1007/s10549-015-3474-4 (2015).

22 Wilkerson, M. D. *et al.* Prediction of lung cancer histological types by RT-qPCR gene expression in FFPE specimens. *The Journal of molecular diagnostics : JMD* **15**, 485-497, doi:10.1016/j.jmoldx.2013.03.007 (2013).

23 Wilkerson, M. D. *et al.* Lung squamous cell carcinoma mRNA expression subtypes are reproducible, clinically important, and correspond to normal cell types. *Clin Cancer Res* **16**, 4864-4875, doi:10.1158/1078-0432.CCR-10-0199 (2010).

24 Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* **436**, 518-524, doi:10.1038/nature03799 (2005).

25 Wilkerson, M. D. *et al.* Differential pathogenesis of lung adenocarcinoma subtypes involving sequence mutations, copy number, chromosomal instability, and methylation. *PloS one* **7**, e36530-e36530, doi:10.1371/journal.pone.0036530 (2012).

26 Chang, H. Y. *et al.* Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proceedings of the National Academy of Sciences of the United States of America* **102**, 3738-3743, doi:10.1073/pnas.0409462102 (2005).

27 Huang, F. *et al.* Identification of candidate molecular markers predicting sensitivity in solid tumors to dasatinib: rationale for patient selection. *Cancer Res* **67**, 2226-2238, doi:10.1158/0008-5472.can-06-3633 (2007).

28 Chi, J.-T. *et al.* Gene expression programs in response to hypoxia: cell type specificity and prognostic significance in human cancers. *PLoS medicine* **3**, e47-e47, doi:10.1371/journal.pmed.0030047 (2006).

29 Liu, R. *et al.* The Prognostic Role of a Gene Signature from Tumorigenic Breast-Cancer Cells. *New England Journal of Medicine* **356**, 217-226, doi:10.1056/NEJMoa063994 (2007).

30 van 't Veer, L. J. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530, doi:10.1038/415530a

https://www.nature.com/articles/415530a#supplementary-information (2002).

31 Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 1160-1167, doi:10.1200/JCO.2008.18.1370 (2009).

32 Oh, D. S. *et al.* Estrogen-Regulated Genes Predict Survival in Hormone Receptor–Positive Breast Cancers. *Journal of Clinical Oncology* **24**, 1656-1664, doi:10.1200/jco.2005.03.2755 (2006).

33 Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506-519, doi:10.1016/j.cell.2015.09.033 (2015).

34 Troester, M. A. *et al.* Gene expression patterns associated with p53 status in breast cancer. *BMC cancer* **6**, 276-276, doi:10.1186/1471-2407-6-276 (2006).

35 Saal, L. H. *et al.* Poor prognosis in carcinoma is associated with a gene expression signature of aberrant PTEN tumor suppressor pathway activity. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 7564-7569, doi:10.1073/pnas.0702507104 (2007).

36 Julka, P. K. *et al.* A phase II study of sequential neoadjuvant gemcitabine plus doxorubicin followed by gemcitabine plus cisplatin in patients with operable breast cancer: prediction of response using molecular profiling. *Br J Cancer* **98**, 1327-1335, doi:10.1038/sj.bjc.6604322 (2008).

37 Cancer Genome Atlas, N. Genomic Classification of Cutaneous Melanoma. *Cell* **161**, 1681-1696, doi:10.1016/j.cell.2015.05.044 (2015).

38 Liu, Y., Hayes, D. N., Nobel, A. & Marron, J. S. Statistical Significance of Clustering for High Dimensional Low Sample Size Data. *Journal of the American Statistical Association* **103** (2008).

39 Tusher, V. G., Tibshirani, R. & Chu, G. Significance analysis of microarrays applied to the ionizing radiation response. *Proc Natl Acad Sci U S A* **98**, 5116-5121 (2001).

40 Benjamini, Y., Drai, D., Elmer, G., Kafkafi, N. & Golani, I. Controlling the false discovery rate in behavior genetics research. *Behavioural brain research* **125**, 279-284, doi:10.1016/s0166-4328(01)00297-2 (2001).

41 Zhou, W., Laird, P. W. & Shen, H. Comprehensive characterization, annotation and innovative use of Infinium DNA methylation BeadChip probes. *Nucleic Acids Res* **45**, e22, doi:10.1093/nar/gkw967 (2017).

42 Zhou, W., Triche, T. J., Laird, P. W. & Shen, H. SeSAMe: reducing artifactual detection of DNA methylation by Infinium BeadChips in genomic deletions. *Nucleic Acids Res* **46**, e123, doi:10.1093/nar/gky691 (2018).

43 Triche, T. J., Weisenberger, D. J., Van Den Berg, D., Laird, P. W. & Siegmund, K. D. Low-level processing of Illumina Infinium DNA Methylation BeadArrays. *Nucleic Acids Res* **41**, e90, doi:10.1093/nar/gkt090 (2013).

44 Davis, C. A. *et al.* The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic Acids Res* **46**, D794-d801, doi:10.1093/nar/gkx1081 (2018).

45 Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995, doi:10.1093/nar/gks1193 (2013).

46 Consortium, E. P. *et al.* Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature* **583**, 699-710, doi:10.1038/s41586-020-2493-4 (2020).

47 Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* **67**, 1 - 48, doi:10.18637/jss.v067.i01 (2015).

48 Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* **57**, 289-300 (1995).

49 Zheng, R. *et al.* Cistrome Data Browser: expanded datasets and new tools for gene regulatory analysis. *Nucleic Acids Res* **47**, D729-D735, doi:10.1093/nar/gky1094 (2019).

50 Sheffield, N. C. & Bock, C. LOLA: enrichment analysis for genomic region sets and regulatory elements in R and Bioconductor. *Bioinformatics* **32**, 587-589, doi:10.1093/bioinformatics/btv612 (2016).

51 Mei, S. *et al.* Cistrome Cancer: A Web Resource for Integrative Gene Regulation Modeling in Cancer. *Cancer Res* **77**, e19-e22, doi:10.1158/0008-5472.CAN-17-0327 (2017).

52 Yu, G., Wang, L. G., Han, Y. & He, Q. Y. clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics : a journal of integrative biology* **16**, 284-287, doi:10.1089/omi.2011.0118 (2012).

53 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248 (2015).

54 Gao, G. F. *et al.* Before and After: Comparison of Legacy and Harmonized TCGA Genomic Data Commons' Data. *Cell Syst* **9**, 24-34 e10, doi:10.1016/j.cels.2019.06.006 (2019).

55 Kelly, B. J. *et al.* Churchill: an ultra-fast, deterministic, highly scalable and balanced parallelization strategy for the discovery of human genetic variation in clinical and population-scale genomics. *Genome Biol* **16**, 6, doi:10.1186/s13059-014-0577-x (2015).

56 Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res* **22**, 1760-1774, doi:10.1101/gr.135350.111 (2012).

57 Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867-868, doi:10.1093/bioinformatics/btx699 (2018).

58 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079, doi:10.1093/bioinformatics/btp352 (2009).

59 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303, doi:gr.107524.110 [pii]

10.1101/gr.107524.110 (2010).

60 Flensburg, C., Sargeant, T., Oshlack, A. & Majewski, I. J. SuperFreq: Integrated mutation detection and clonal tracking in cancer. *PLoS Comput Biol* **16**, e1007603, doi:10.1371/journal.pcbi.1007603 (2020).

61 Koboldt, D. C., Larson, D. E. & Wilson, R. K. Using VarScan 2 for Germline Variant Calling and Somatic Mutation Detection. *Curr Protoc Bioinformatics* **44**, 15.14.11-17, doi:10.1002/0471250953.bi1504s44 (2013).

62 Wright, S. The Interpretation of Population Structure by F-Statistics with Special Regard to Systems of Mating. *Evolution* **19**, 395-420, doi:10.2307/2406450 (1965).

63 Shannon, C. E. The mathematical theory of communication. 1963. *MD Comput* **14**, 306-317 (1997).

64 Hundal, J. *et al.* pVACtools: A Computational Toolkit to Identify and Visualize Cancer Neoantigens. *Cancer Immunol Res* **8**, 409-420, doi:10.1158/2326-6066.CIR-19-0401 (2020).