Phased Whole-Genome Detection and Analysis of Structural Variants of Invasive Ductal and Lobular Breast Cancer Cell Lines

By

Erin Fey

Thesis

Submitted to the Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

MASTERS

in

Pathology

December 14, 2019

Nashville, Tennessee

Approved:

Thomas Stricker, M.D., Ph.D.

Andries Zijlstra, Ph.D.

Justin Balko, Pharm. D., Ph.D.

Deborah Lannigan, Ph.D.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Breast Cancer

Worldwide, breast cancer is the most commonly diagnosed cancer in women, other than nonmelanoma skin cancer.[1] In 2017 more than 250,000 breast cancer cases were diagnosed in the US and over 40,000 breast cancer related deaths occurred.[2] Breast cancers exhibit a large range of morphological features, immunohistochemial profiles, and histological subtypes that can dictate their clinical course of treatment and outcome. Breast cancers can be subclassified based on histologic criteria (ductal versus lobular) as well as molecular profiling (estrogen receptor and progesterone receptor expression, HER2 amplification or triple negative). Breast cancers can also be subclassified based on gene expression profiles, and these intrinsic subtypes incompletely overlap with molecular subtypes.

Approximately 70% of all breast cancers are driven by estrogen receptor-α (ERα).[3] Hormone receptor positivity remains the central feature of this disease. ERα is a steroid hormone receptor and a transcription factor. When it binds estrogen, ERα activates oncogenic pathways in breast cancer cells.[4] ERα positive tumors are treated with anti-estrogen therapy. ERα targeted therapies have been extremely successful in improving outcomes.[5] However ER+ breast cancers are heterogeneous and exhibit significant variability in biological behavior, response to therapy and outcome.[6] Although endocrine therapy has been successful in treating breast cancers, endocrine resistance, both de novo and acquired, remains a critical dilemma. Indeed, around 30-50% of early breast cancer patients will relapse due to acquired resistance.[7] Understanding the

molecular mechanisms underlying the diverse behavior of these tumors is critical to tailoring current therapies and developing new ones.

Histological Subtypes of Breast Cancer

Breast cancer is a histologic diagnosis made according to pathological criteria. The most common breast cancer histology is invasive ductal carcinoma (IDC), accounting for 70-80% of breast cancers. Invasive lobular carcinoma (ILC) constitutes around 10-15% of breast cancer cases.[2] Various types of rare histologies as well as mixed ductal/lobular make up the remainder of breast cancers.[1] Originally, IDCs were thought to develop from the breast ducts, while ILCs developed from the lobules. However, we now know that both IDC and ILC arise from the same segment of the terminal duct lobular unit.[4] Although these subtypes arise from the same structure, they differ in epidemiology, genetic signatures and histology. Compared to IDC, ILC is difficult to detect on mammography and tends to show a worse long-term outcome with a higher incidence of metastasis, recurrence and breast cancer mortality.[8] IDCs are a heterogenous group of tumors that can appear as diffuse sheets, nests or singly distributed cells with different amounts of ductal differentiation.[4] ILCs are characterized morphologically as single-file, small, round, discohesive cells. This phenotype is a consequence of the deregulation of cell-cell adhesion properties caused by the loss of E-cadherin expression. Loss of E-cadherin is found in about 90% of ILCs and thus is its main genomic feature.[9] Currently IDC and ILC tumors are treated the same way. Recognizing that biological heterogeneity underlies histological heterogeneity, several studies have been aimed to focus on the molecular characterization of IDC and ILC. Studies have emerged to determine the genomic landscape of these two subtypes. Cirello and colleagues performed a comprehensive analysis of 817 breast tumor samples (127

ILCs, 490 IDCs and 88 mixed) [Ciriello et al. 2015] and Desmedt and colleagues performed a comprehensive analysis of 417 ILC tumors [Desmedt et al. 2016]. Mutations targeting PTEN, TBX3 and FOXA1 as well as activation of AKT are enriched in ILC compared to IDC.[10] Mutations in FOXA1 correlated with increased FOXA1 expression in ILC while GATA3 mutations correlated with increased GATA3 expression in IDC. Additionally, ESR1 copy number gains were more frequent in ILC than IDC. These gains were associated with higher ESR1 mRNA levels as well as mRNA expression of TFF1, a canonical ESR1 transcriptional target.[41] These studies clearly showed that ILC has distinct genomic features compared to IDC. Of importance, the distribution of hotspot mutations in genes such as FOXA1 and GATA3 in ILC and IDC may have implications in response to therapies. While these studies identified functional genomic characteristics in IDC and ILC tumors, higher-order structural features of their genomes have yet to be characterized. These distinct molecular portraits between the histological subtypes of breast cancer highlight the need for individualized therapies based on histology.

Copy Number Alterations/Structural Variations

One of the hallmarks of cancer cells is genomic instability.[11] Genomic instability generates mutations and large-scale structural variations like chromosomal translocations and copy number alterations that can drive tumor progression. Genome instability plays a critical role in cancer initiation, progression, evolution and drug resistance through reduced apoptosis, unchecked proliferation, increased motility and angiogenesis.[12] Copy number alterations (CNAs) are gains or losses in copies of DNA segments and are present in many types of cancer.[13] CNAs affect a greater fraction of the genome than single nucleotide polymorphisms (SNPs).

Furthermore, it is reported that 85% of the variation in gene expression of breast tumors are due to somatic CNAs.[14]

The first TCGA breast cancer study reported on 466 breast cancer tumors on six different technology platforms, one being DNA copy number analysis.[10] They identified copy number alterations that are associated with molecular subtypes. As previously mentioned, Cirello and colleagues used a larger cohort and also performed copy number analysis and found that the frequency of copy-number alterations that are known breast cancer gains and losses, differed significantly between IDC and ILC. Copy number alterations were identified using Affymetrix SNP arrays to determine CNA. Affymetrix arrays use probes to detect SNPs as well as non-polymorphic probes to detect CNAs.[15] However, the CNA data obtained from these two studies lack resolution and sensitivity. Non-linked read CNAs and WES are other methods of finding CNAs. However, these methods are inferior to 10X linked reads because many structural variants are significantly longer than the DNA libraries produced by these technologies, whose insert lengths are about 300-500 nucleotides.[16] In addition, such reads are too short for accurate de novo genome assembly.

There are hundreds of regions of the genome that are recurrently amplified and deleted, and most of these do not encompass known oncogenes or tumor suppressor genes. Identification of recurrent CNAs, which are reported to have a strong association with clinical phenotypes, has resulted in new therapeutic options.[17] Therefore, CNAs in breast cancer patients could be regarded as potential biomarkers, presenting the opportunity for new therapeutics. Determining the genes that are targeted by CNAs will benefit the mechanism by which those CNAs arise as well as the positive and negative effects on gene expression.

10X Phased Sequencing

Whole genome sequencing has produced tens of thousands of genomes that are collections of short-read sequences aligned to the composite reference human genome sequence. It is cost-effective, high-throughput and accurately calls bases but it fails to reliably call structural variants, assess variation across the entire genome and reconstruct long-range haplotypes. Most genome analyses are performed with short reads, resulting in analyses of small variants over nonrepetitive parts of the genome.[18] Structural variants, especially those larger than a few thousand bases or those that are in repetitive elements, are almost impossible to resolve with short-read sequencing. Thus, we are underrepresenting the amount of structural variation in the genome when using these sequencing approaches.

10X is a synthetic long-read technology that works by using as little as 1 ng of high molecular weight DNA that is partitioned (100kb) into micelles known as Gel-bead in EMulsions (GEMs). Each GEM contains approximately 0.3x genome copies and a unique barcode.[19] The long pieces of DNA in each droplet are fragmented and barcoded. These fragments are then used for library building and sequencing. Each long piece of DNA in a micelle has the same barcode, and so must have been close together in space. Thus, after sequencing, the barcoded short-reads can be assembled into continuous sequences through their unique barcode, known as linked reads. This technology is able to reconstruct long-range information from short-reads, unlike current whole-genome sequencing methods

Linked reads allow mapping to 38Mb of sequence not accessible to short-reads, making 423 difficult to sequence genes accessible to NGS.[8] 10X can give more information about structural variation in cancer cells and resolve maternal and paternal haplotypes. [20] Furthermore, it has low input requirements and error rates.

My Project

We wanted to utilize 10X genomics as our technology to understand the complete landscape of structural variant changes of invasive ductal and invasive lobular carcinomas. Structural variants are key to cancer development, and improved identification of structural variants will lead to new insights into molecular types in cancer. Advantages to using cell lines for these experiments are that they are relatively genomically pure, such that there is no contamination from normal cell infiltrate, they are cell characterized, and it is easy to get large pieces of DNA from them. MCF7 and T47D are two of the most widely used ER+ IDC cell lines, with many more extensively studied in the literature. However, very few ER+ ILC cell lines have been reported in the literature, MDA-MB-134 (MM134) and SUM44PE being the most widely used.

From our 10X data we can easily assemble genomes, find more structural variants (SNVs, deletions, amplifications and translocations), access areas of the genome previously inaccessible and resolve haplotypes from breast cancer histological subtypes. Furthermore, we can use 10X data to better understand ER regulation. Our 10X sequencing data coupled with ChIP-seq ER data allows us to learn more about how ER is affected by structural variants. Comparing this data with RNA-seq we can also gain more information as to how ER is driving gene expression.

In conclusion we sequenced our four cell lines, MCF7, T47D, MM134 and SUM44, using 10X linked read sequencing and were able to identify structural variants in great detail. We characterized amplifications, deletions and translocations found in each of the four cell lines as well as structural variants the cell lines share, specifically the histological subtypes. We also

further characterized structural variants that were already known in these cell lines. What we have found is that structural variation in cancer cell lines is diverse. Each cell line has hundred of unique structural variants and it will be pertinent to discover what structural variants are functionally relevant.

CHAPTER 2

ANALYSIS

Structural Variant Analysis

10X sequencing technology uses the Lariat aligner through the Long Ranger pipeline to align barcoded linked reads. All the linked reads for a single barcode are aligned simultaneously, with the prior knowledge that the reads arise from a small number of long (10kb-200kb) molecules.[21] The large-scale structural variant caller looks for distant pairs of loci in the genome that share many more barcodes than would be expected by chance. Any overlap indicated that the two loci that are very distant in the reference sequence are close in the sample and generates a candidate structural variant. Candidate structural variants are then refined by comparing the layout of reads and barcodes around the event and the patterns expected in deletions, inversions, duplications and translocations to identify the type of structural variants and also find the maximum-likelihood of breakpoints.[21]

Files generated from the Long Ranger pipeline are then available using Loupe genome browser, which allows for easy visualization of the data. Included in this browser is visualization of the barcode overlap evidence for large-scale structural variants called by Long Ranger. For each variant, the structural variant list provides a quality score for the variant, the locations of the two breakpoints (chromosome and position), a list of genes that are close to the breakpoints and the distance between the breakpoints. The quality score is a log-likelihood score comparing that there is a structural variant between two loci or that the observed barcode overlap between two

loci was generated by chance. The higher the score, the stronger the evidence is that there is a structural variant.

To build a detailed map of breast cancer copy number changes, we used 10X whole genome sequencing on four commonly studied breast cancer cell lines: 2 ductal cell lines (MCF7, T47D) and 2 lobular cell lines (SUM44, MM134). 10X identified 1,059 structural variants in the MCF7 cell line, composed of 575 amplifications, 311 deletions and 173 translocations. Altogether, 32.7% of the MCF7 genome was involved in copy number changes, with 12.6% of the genome involved in amplifications and 20.1% of the genome involved in deletions. 775 structural variants were identified in the T47D cell line, composed of 309 amplifications 423 deletions and 43 translocations. Altogether, 66.3% of the T47D genome was involved in copy number changes, with 27.9% of the genome involved in amplifications and 38.4% of the genome involved in deletions. 463 structural variants were identified in SUM44, composed of 277 amplifications, 145 deletions and 41 translocations. Altogether, 35.3% of the SUM44 genome was involved in copy number changes, with 5.9% of the genome involved in amplifications and 29.4% of the genome involved in deletions. 182 structural variants were identified in MM134, composed of 70 amplifications, 64 deletions and 48 translocations. (Figure 1 and Table 1). Altogether, 15.6% of the MM134 genome was involved in copy number changes, with 4.8% of the genome involved in amplifications and 10.8% of the genome involved in deletions.

We observed that our ductal cell lines have more amplifications than our lobular cell lines. MCF7 has the largest number of translocations, while T47D has the largest fraction of the genome affected by copy number changes. Lobular cell lines have a smaller fraction of the genome affected by copy number variants, compared to lobular cell lines. Table 1 lists the

average size of amplifications and deletions and we observe that although a small fraction of the genome is involved in copy number changes, there is a slightly larger average deletion size in lobular cell lines. We next determined how many genes are included in amplifications and deletions. For this analysis we used any overlapping sequence of an amplification or deletion with the gene to count that gene. We found that on average more genes were involved in amplifications than deletions. MCF7 had 5737 genes involved in amplifications and 46 genes involved in deletions. T47D had 13182 genes involved in amplifications and 222 genes involved in deletions. MM134 had 1862 genes involved in amplifications and 18 genes involved in deletions. SUM44 had 2521 genes involved in amplifications and 46 genes involved in deletions. This is interesting because our average amplification distance is smaller for all of the cell lines than the average deletion distance, yet the amplifications seem to be affecting more genes. However, T47D and MM134 have a much greater average amplification distance and have a much greater number of genes involved, which makes sense. Determining genes involved in translocations we identified known genes at the start and end section of the chromosome. MCF7 has potentially 120 gene fusions, 26 potential gene fusions in T47D, 38 potential gene fusions in MM134 and 35 potential gene fusions in SUM44.

Combining all our amplifications, deletions and translocations of all four cell lines we have 2,447 structural variants. Of those structural variants 629 (25%) occur in 2 cell lines, 154 (6%) occur in 3 cell lines and 14 (0.6%) occur in all four cell lines. Of the variants that occur in all four cell lines, half occur in chromosome 11. When we categorize for histological subtype, IDCs have a combined 1,834 structural variants and ILCs 645. Within the two IDC cell lines 316 variants occur in both MCF7 and T47D (17%). Strikingly, only 10 variants occur in both MM134 and SUM44 (1.5%).

Additionally we constructed circos plots of our structural variant data to better visualize the structural variants within the genome (Fig. 1). Using these methods we identified patterns of amplifications, deletions, and translocations. Not only do ductal cell lines contain more amplifications and deletions than the lobular cell lines but we see that the chromosomal regions that are being amplified for deleted appear to be the same in the two cell lines.

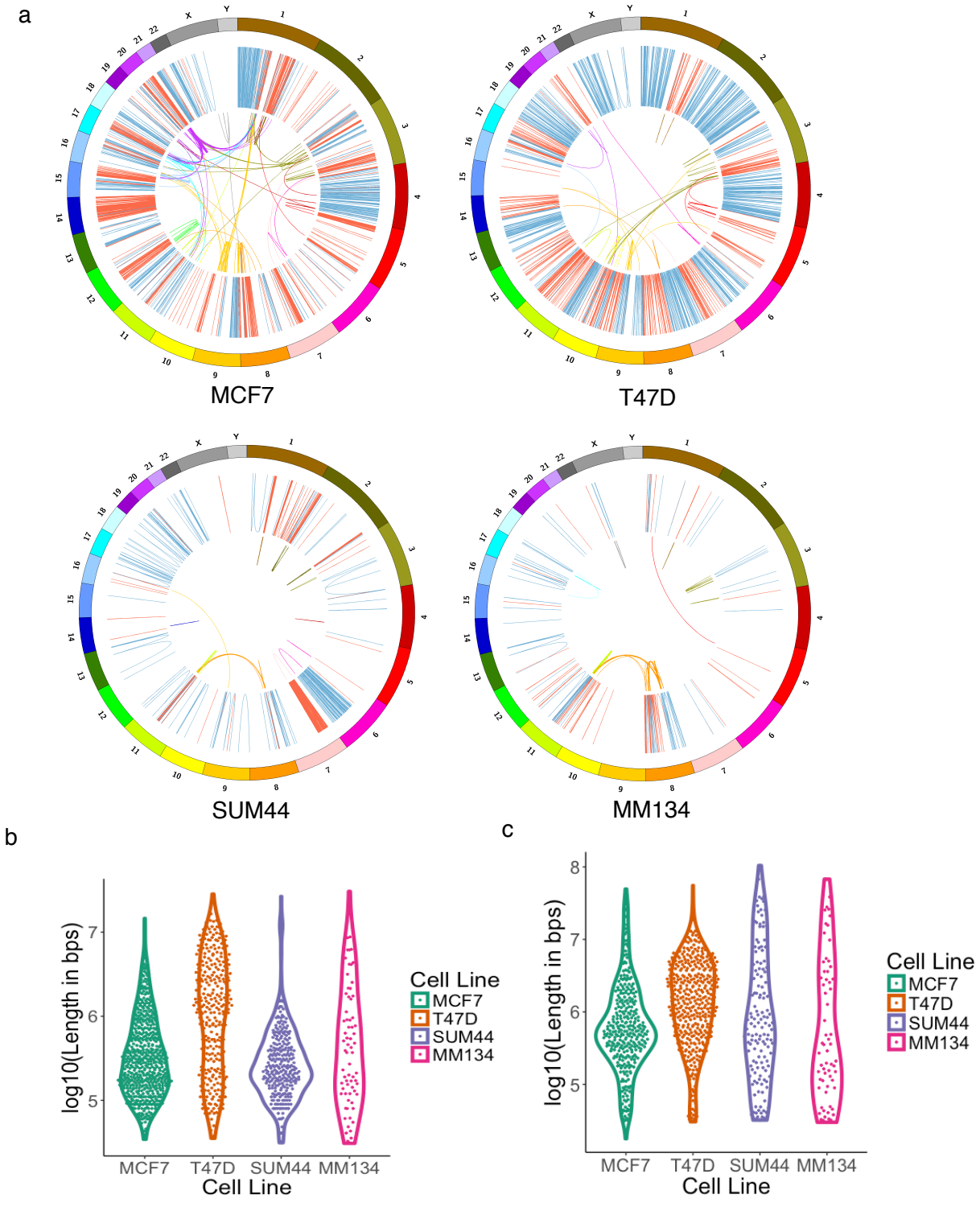


Fig.1 | Structural variants found in cancer genomes. a. Circos plots of all four cancer genomes. The tracks from outer to innercircles are chromosome coordinates, duplications (red) and deletions (blue), and translocations (The color of the link is set to the 2nd chromosome in the link coordinate) b. Violin plot of number and size distribution of amplifications in each cell line c. Violin plot of deletions

Table 1 | Average distance (bps) and number of structural variants in cancer cell line genomes

	Average amplification distance	Number of amplifications	Average deletion distance	Number of deletions	Number of translocations
MCF7	651579	575	1893122	311	173
T47D	2696910	309	2743016	423	43
SUM44	628842	277	5584731	145	41
MM134	1676572	70	4897521	64	48

Gene Expression Analysis

To identify transcriptomic alterations, we performed RNA-seq on our four cell lines. A principal component analysis (PCA) was performed on the duplicated biological replicates to determine the level of gene expression similarity (Figure 2a). We can observe that the replicates of each cell line cluster closely together which is what we would expect. Additionally each cell line has unique variation, but we can note that MM134 and SUM44 appear to cluster closer together.

We used DESeq2 to identify differentially expressed genes (DEGs) between our cell lines. Our negative binomial linearized model compares differential expression of genes between histological subtypes. Applying this model, we used an interaction term for histological subtype and were able to get a list of DEGs between subtypes to determine if ER is regulating different genes in ductal vs. lobular. We found 6,978 DEG (FDR 0.05) in our ductal vs. lobular cell lines. We took these genes and ran iRegulon through Cytoscape to identify enriched motifs and transcription factors in our gene set. Our top hits included BHLHE40, ATF4 and NFKB1 (Figure 2b). BHLHE40 is a transcription factor that is directly activated by HIF1A under hypoxia and has been shown to confer a pro-survival and pro-metastatic phenotype to breast cancer cells.[44] ATF4 is a transcription factor that upregulates genes involved in amino acid transport, glutathione biosynthesis and the antioxidative stress response.[45] NFKB1 is a subunit of NKFB which has been shown to be a suppressor of aging, inflammation and cancer.[46]

We next wanted to take a closer look at our amplifications and deletions and determine if they were affecting gene expression. Figure 2c is an oncoplot of specific genes that are amplified or deleted in at least two cell lines. While previous analysis identified much higher numbers for structural variants found in multiple cell lines, this analysis only looked at structural variants that

occurred on a known gene as well as amplified the entire gene. Two genes that stuck out to us were LRP1B and DPYD. Both of these genes have been shown to have an effect on chemotherapy treatment. LR1PB plays a role in the drug endocytosis of the chemotherapeutic agent, liposomal doxirubicin.[42] It has been shown that deletion or mutation of LR1PB is associated with acquired chemotherapy resistance. LR1PB is deleted in three of our cell lines. The DPYD gene encodes DPD, which is an enzyme that catalyzes fluorouracil metabolism.[43] Fluorouracil is a chemotherapy agent used to treat many different cancers. Deletion, or loss of function mutations of DPYD may not be able to metabolize fluorouracil at normal rates leading to potentially life-threatening fluorouracil toxicity. T47D has a DPYD deletion and SUM44 has DPYD amplification.

To get a better idea of how these amplifications and deletions were affecting gene expression, we subset either amplifications or deletions for each cell line. We then made a heat map of Z-scores of gene expression. Between all cell lines there was no clear observation that when a gene was amplified it necessarily had higher gene expression levels. For example in Figure 2d, we have a list of genes that are amplified in MCF7, however not all of those genes have higher gene expression levels (indicated by green boxes). However it is important to note in this heat map, that when genes are upregulated in MCF7 they tend to be upregulated in T47D but downregulated in the lobular cell lines (and vice versa). This is representative of our other cell lines. Additionally, we took the list of genes amplified or deleted in each cell line and plotted box plots of gene expression. In Figure 2e dots in red represent the genes amplified in MCF7 compared to all other genes. We see that most dots fall in the range of gene expression of all other and only a few express higher levels. This is representative of our other cell lines. We also determined how many of our differentially expressed genes were amplified or deleted.

Comparing all of our amplified gene names with our ductal vs. lobular DEG list we only found 36 genes that are both amplified and differentially expressed. We found 155 genes that had some deletion in them and were differentially expressed.

Next we wanted to generate DEG for each individual cell line, such that one cell line vs. all the others, to determine what DEGs are structural variants. Again we used DESeq2 but this time we included an interaction term for cell line. We found 5,530 genes differentially expressed in MCF7 cells wherein 148 had some structural variant. 3,904 genes were differentially expressed in T47D and 114 had some structural variant. We found much lower numbers for DEGs in our lobular cell lines. MM134 had 258 DEGs wherein only 10 were a structural variant and SUM44 had 272 DEGs and only 5 had some structural variant. We did a comparison to determine if we had any overlap in structural variants that were differentially expressed between cell lines and we only found a very small number of genes that were in two cell lines (Figure 2f).

We also did an analysis on ER binding sites in our amplified regions. Using ChIP-seq data we had locations of ER binding in each of our cell lines. We coupled this with our amplification data only selecting for amplifications smaller than 500,000 bps. We made this cutoff so we could determine specific focal amplifications of ER binding sites and anything bigger than 500,000 are unfocused. Using bedtools intersect we were able to determine the number of times ER binding sites intersected with an amplification (Table 2). Most striking is the number of intersections between ER binding and amplifications in MM134. There were only 39 amplifications identified under 500,000 bps and of those 33 intersected with ER binding sites. The other cell lines had around 25-50% intersection. MM134 has a high average amplification distance (1676572 bps) so the cutoff removed many of the larger amplifications (Table 1). However it is interesting that the small amplifications MM134 does have are almost always

intersecting with an ER binding site. While amplifications are not always affecting the gene expression of the gene that is amplified, they could be having an affect on ER levels and regulation.

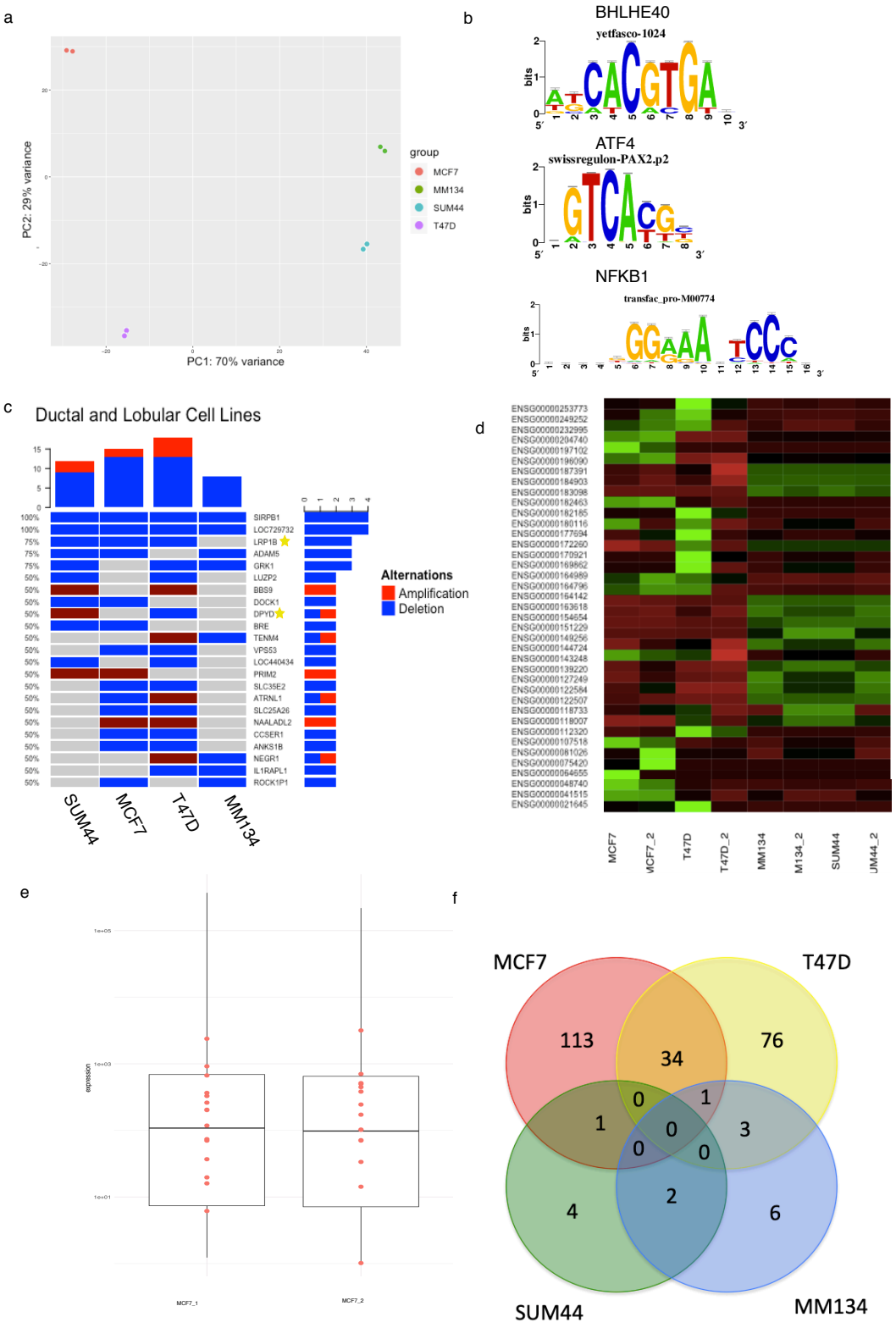


Fig. 2 | Gene expression analysis a. PCA from RNA-seq data of all the genes of all 4 cell lines and replicates. b. Top motifs in our list of differentially expressed genes, BHLHE40, ATF4 and NKKB1. c. Oncoplot of amplifications and deletions identified using phased genomes. This oncoplot is only representative of genes that have amplifications or deletions in at least two cell lines. Amplifications of only the whole gene were taken into account for this analysis. d. Heat map of MCF7 amplified genes e. Box plot of gene expression of all genes. Dots in red represent the list of genes that are amplified in MCF7 (as listed in d). f. Venn diagram comparing DEGs that are also structural variants for each cell line.

Table 2 | Number of ER binding sites and amplifications and intersection between them

	Number of ER binding sites	Number of amps < 500,000bps	Number of intersections
MCF7	5895	390	169
T47D	5676	105	28
SUM44	1718	216	55
MM134	5670	39	33

Characterization of CCDC170-ESR1 Fusion

Gene fusions resulting from genomic rearrangements are important drivers for cancer initiation and progression. Both de novo and acquired resistance to endocrine therapy for ER+ breast cancers remains a significant clinical challenge. Recurrent point mutations around the ligand binding domain of estrogen receptor alpha gene (ESR1) have been found in up to 40% of post-treatment metastatic breast cancer patients, however these mutations fail to explain most cases of endocrine resistance.[23] Evidence now suggests that ESR1 fusions are another class of mutations associated with endocrine resistance.[23] Many ESR1 fusions have been identified in breast cancer, but their role in breast cancer is not completely understood.

CCDC170-ESR1 fusion involves the first two non-coding exons of ESR1 fused to various C-termini sequences from the coiled-coil domain containing protein, CCDC170.[24] This fusion generates truncated forms of CCDC170 proteins that, when introduced into ER+ breast cancer cells, reduced endocrine sensitivity.[25] Another study identified ESR1-CCDC170 as a fusion that occurred with endocrine therapy resistance after letrozole treatment.[26] ESR1-CCDC170 has previously been identified in the MCF7 cell line, as well as about 3.5% of ER+ breast cancer cases in the TCGA.[25] Additionally, previous studies have shown that an enhancer region of androgen receptor is a driver in castrate-resistant prostate cancer and androgen receptor is frequently amplified.[47,48] This led us to hypothesize that this amplification in MCF7 cells could be driving ER.

Using our 10X data, we identified this fusion in MCF7. Interestingly, the increased resolution of the 10X data allowed identification of an amplified region within the CCDC170-ESR1 fusion (Figure 3a). Utilizing UCSC Genome Browser, we observed that this region is essentially the ESR1 promoter. Interestingly, this region was amplified at least three times, with

three sets of independent breakpoints. One of these amplifications was an inversion, which creates the CCDC170-ESR1 fusion. Given this information we wanted to investigate whether this amplified region could be driving gene expression of ESR1. In TCGA, 14 breast cancer tumors from 660 ER+ primary breast samples are identified as having this ESR1 fusion (2.1%). Within those 14 breast tumors, two of the fusions occur out-of-frame while the rest are in the 5'UTR-coding region. We compared ESR1 mRNA levels of these fusion positive tumors against fusion negative numbers (Figure 3e). We observed that tumors harboring the CCDC170-ESR1 fusion had slightly higher ESR1 levels than tumors that did not have the fusion.

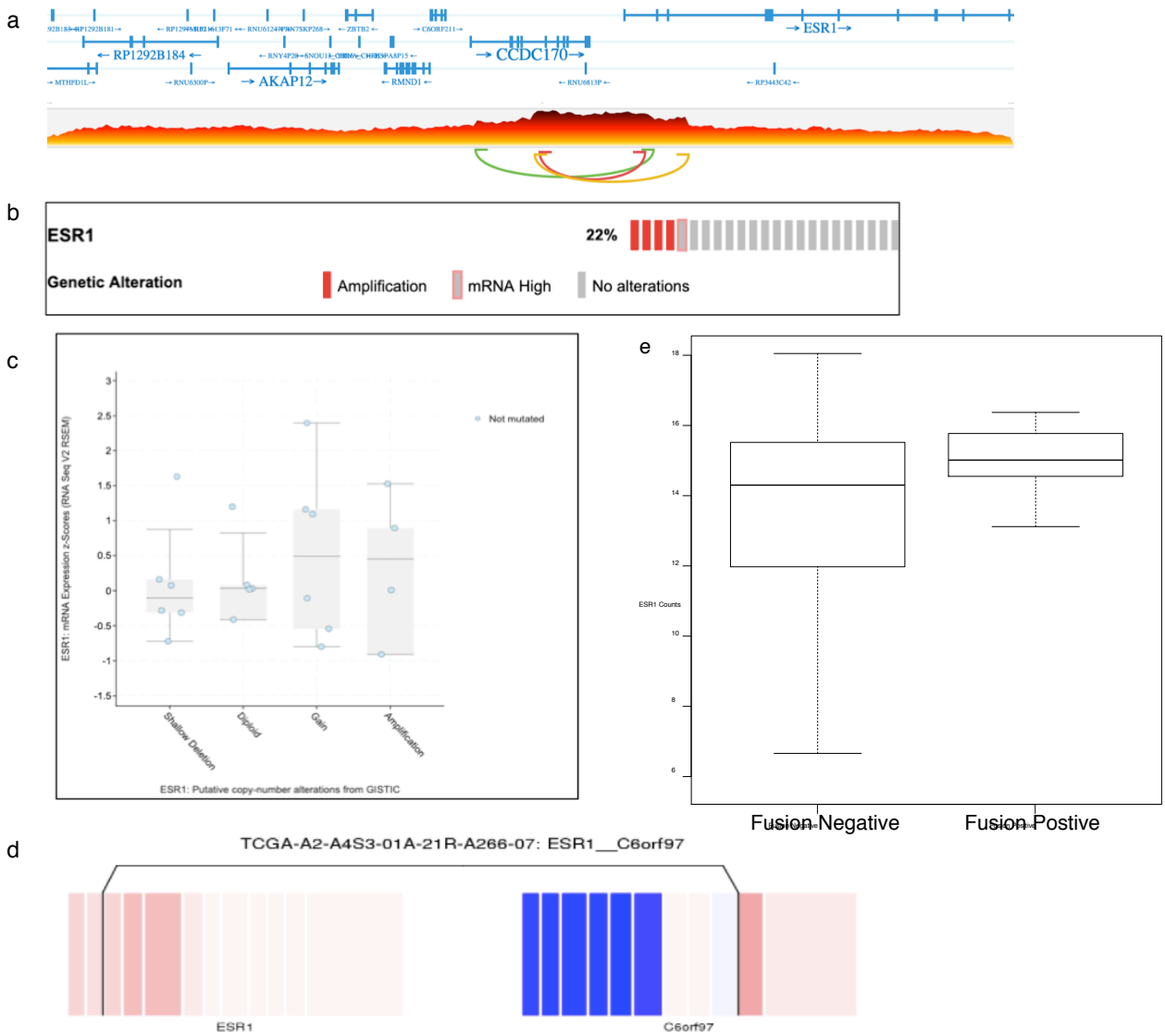


Fig. 3 | Characterization of CCDC170-ESR1 fusion in MCF7 cells. a. Track taken from Loupe brower of 10X phased MCF7 sequencing data observing the CCDC170-ESR1 junction. The taller red sections identifies an amplified region while the green, red, and yellow arcs are where the translocation occurs. b. 23 breast cancer tumors in TCGA were identified having this fusion. This oncoprint show that within those 23 tumors 22% had amplified ESR1. c. Box plot of ESR1 mRNA levels taken from TCGA samples. Compared to CNA, ESR1 mRNA levels are not much difference in expression so this region may not be having an effect on gene expression d. Exon expression plot for fused genes of a TCGA sample with the CCDC170-ESR1 fusion. Expression was normalized across all exons; blue = lowest expression, red = highest expression. Line indicates where genes are connected. C6orf97 is another name for CCDC170. e. ESR1 read counts of TCGA samples. Fusion negative are 546 breast tumors without the CCDC170-ESR1 fusion, fusion positive are the 14 breast tumors that do have this fusion.

Translocations Between Chromosomes 8 and 11 in Invasive Lobular Carcinomas

Depending on the chromosome breakpoint, a translocation can result in the misregulation of normal gene function or the fusion of genes. In many cases, these gene rearrangements are considered to be the primary cause of various cancers. Studies suggest that around 20% of all cancers are caused by chromosomal translocations.[27] For example, the translocation of chromosomes 9 and 22 causes chronic myelogenous leukemia (CML).[28] Chromosomal translocations like these are used as diagnostic markers for cancer and its therapeutics.

We discovered a pattern of translocations in our two lobular cell lines, MM134 and SUM44, between chromosomes 8 and 11 (Figure 4a). While all translocations involved different genes they were all within a few hundred kilobases from each other. Chromosome 8 is 145 Mbs in length and the translocations occurred in a 42.3 Mb region. Chromosome 11 is 135 Mbs in length and the translocations occurred in a 37.5 Mb region (Figure 4b). To determine if this pattern was evident in breast cancer samples, we searched for this translocations in 550 breast cancer cases in TCGA. Of those 550 cases, 9 included 8 to 11 translocations in the region of the cell line translocations (Figure 4a). Two of these samples had two translocations in this region. Although ILC tumors composed only 15.5% of samples, they accounted for 44.4% of 8:11 translocation.

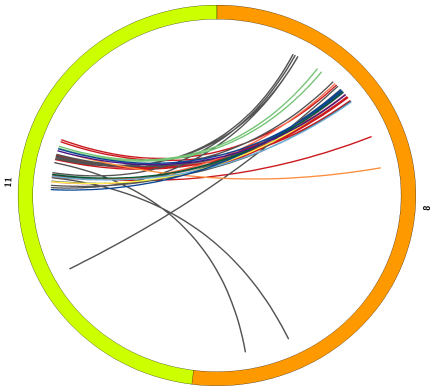
Using gene expression levels from our RNA-seq data we were able to determine the gene expression of genes in this region. Combining a list of genes from chromosome 8 and chromosome 11 we were able to cluster genes according to expression and cell line using the R package, Heatmap.2 (Figure 4e). It is evident within the ~700 genes in these regions that there is a difference in gene expression between our ductal and lobular cell lines. When many genes are upregulated in the lobular cell lines (green) they are downregulated in our ductal cell lines (red).

Therefore, this region of translocation could be playing a role in gene expression in a histological subtype specific way.

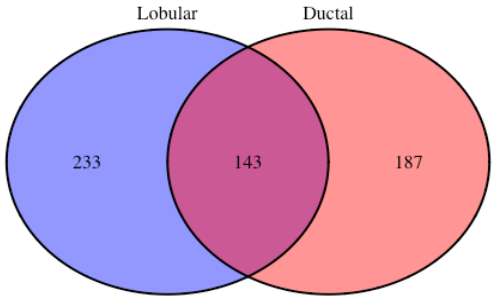
We next hypothesized that ERα binding may play an important role in gene regulation in this region. Using cell line specific ERα ChIP-seq data, we identified 243 ERα binding sites in MM134 and 133 in SUM44. In the same 8:11 region, we see similar ERα binding site in the to ductal cell lines (191 in MCF7 and 139 in T47D). However, 233 ERα binding sites are unique to the lobular subtype (Figure 4c). To determine if there are more ER binding sites in this region than you would expect by chance, we sampled the same Mb size in random regions of either 8 or 11 chromosomes 1000 times and determined how many ER binding sites are in those regions. Bedtools shuffle allowed us to generate a list of random genomic regions so we could use bedtools intersect to overlap this with ER ChIP. we took this list and found that there were 32 times that the number of shuffled intersections exceeded my observed out of 1000 trials. Therefore, our p-value = 0.0359 indicated that the translocation regions of these chromosomes have statistically more ER binding sites then random regions of the chromosomes.

Interestingly a t(8;11) translocation between DOC4 and NRG1, was identified in the breast cancer cell line MDA-MB-175 (ER+, IDC).[27] This translocation results in Y-heregulin which is an autocrine factor under the regulation of the DOC4 promotor, that is implicated in the proliferation of breast cancer cells. While these specific genes are not involved in any of the translocations that we have identified, it is interesting to note that t(8:11) translocations are found in other breast cancer cell lines and patient samples.

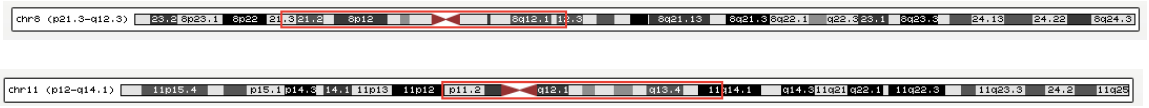
a



c



b



d

Cell Line/TCGA Sample	Histological Subtype	Chromosome 8 Gene	Chromosome 11 Gene
SUM44	ILC	PRKRIR	ANK1
SUM44	ILC	UVRAG	ZMAT4
SUM44	ILC	MTL5	ZMAT4
MM134	ILC	RP112588J	-
MM134	ILC	KLH35	-
MM134	ILC	DMTN	AP002387
MM134	ILC	IGHMBP2	-
MM134	ILC	PAK1	-
E9.A228	IDC	ANO1	UNC5D
A2.A3KD	ILC	ASPH	RSF1
E2.A15R	IDC	CPT1A	ADAM2
A2.A0ES	ILC	GTF2E2	NARS2
AR.A2LK	ILC	IKBKB	FADD
A2.A0ES	ILC	KIF13B	ODZ4
D8.A1XF	IDC	PLEKHA2	SHANK2
A0.A1KT	IDC	RAB11FIP1	ZDHHC24
XX.A899	ILC	RSF1	ADAM2
AC.A2BM	IDC	SHANK2	TACC1
A0.A1KT	IDC	WHSC1L1	GAB2

e

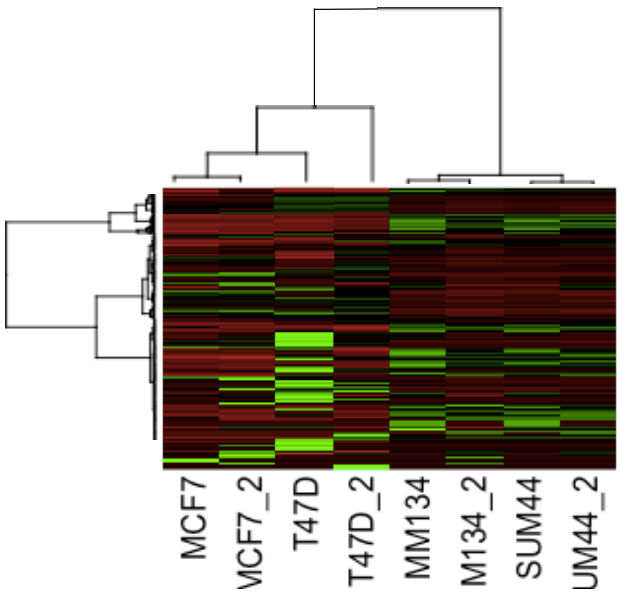


Fig. 6 Characterization of 8-11 translocation in lobular cell lines. a. Circos plot of chromosome 8 and chromosome 11 translocations. Each color arc represents either MM134 cell line (gray), SUM44 cell line (red) or one of the 9 TCGA breast cancer samples. b. Chromosome tracks take from UCSC Genome brower to highlight the regions the translocations take place. c. ER binding sites in the 8-11 chromosome regions for lobular and ductal cell lines. d. Table identifying the genes involved in the translocations.

FGFR1 Amplification in SUM44 Lobular Cell Line

Fibroblast growth factors and their receptors (FGFRs) are involved in different physiologic processes and play important roles in cancer proliferation, survival, differentiation and apoptosis.[29] FGFR alterations have been found in 7.1% of cancers, with the majority being gene amplifications (66%). In breast cancer, FGFR1 is amplified in about 8.7% of patients.[30] FGFR1 gene amplifications are associated with de novo endocrine resistance.[31] Furthermore, tumors harboring FGFR1 amplification displayed a worse distant metastasis-free survival.

Studies have shown that FGFR1 is over-expressed in the MM134 and SUM44 cell line and through viability assays that SUM44 is resistant to 4-hydroxytamoxifen.[31,32] This resistance was reversed when cells were treated with siRNA against FGFR1. Using our 10X data we wanted to better characterize this amplification in the ILC cell lines. Using our RNA-seq data we confirmed that FGFR1 is overexpressed in both MM134 and SUM44 (Figure 5a). We took the genes located on the 8p11.2-p12 amplicon of ILCs where FGFR1 is located at 8p11.23. As we can see many genes in this amplicon of MM134 and SUM44 have high levels of expression compared to our IDC cell lines.

We next wanted to view FGFR1 using the Loupe browser through 10X Genomics. This allows us to visualize our linked-read data. The structural variants view of this application allows us to look for the FGFR1 amplification. We located FGFR1 and confirmed a highly amplified region in both SUM44 and MM134 (Figure 5b,c). Each position in the matrix corresponds to a pair of loci from the two axes. The darker the color the greater number of barcodes that were observed in reads from both loci, therefore, the dark red regions are indicative of areas of amplification. Additionally, the chromosome track above the matrix highlights breakpoints within the region you are looking at. We see numerous breakpoints at the end of the

amplification in both SUM44 and MM134. However, the amplified region is completely missing in our ductal cell lines (Figure 5d,e) and breakpoints are missing in MCF7.

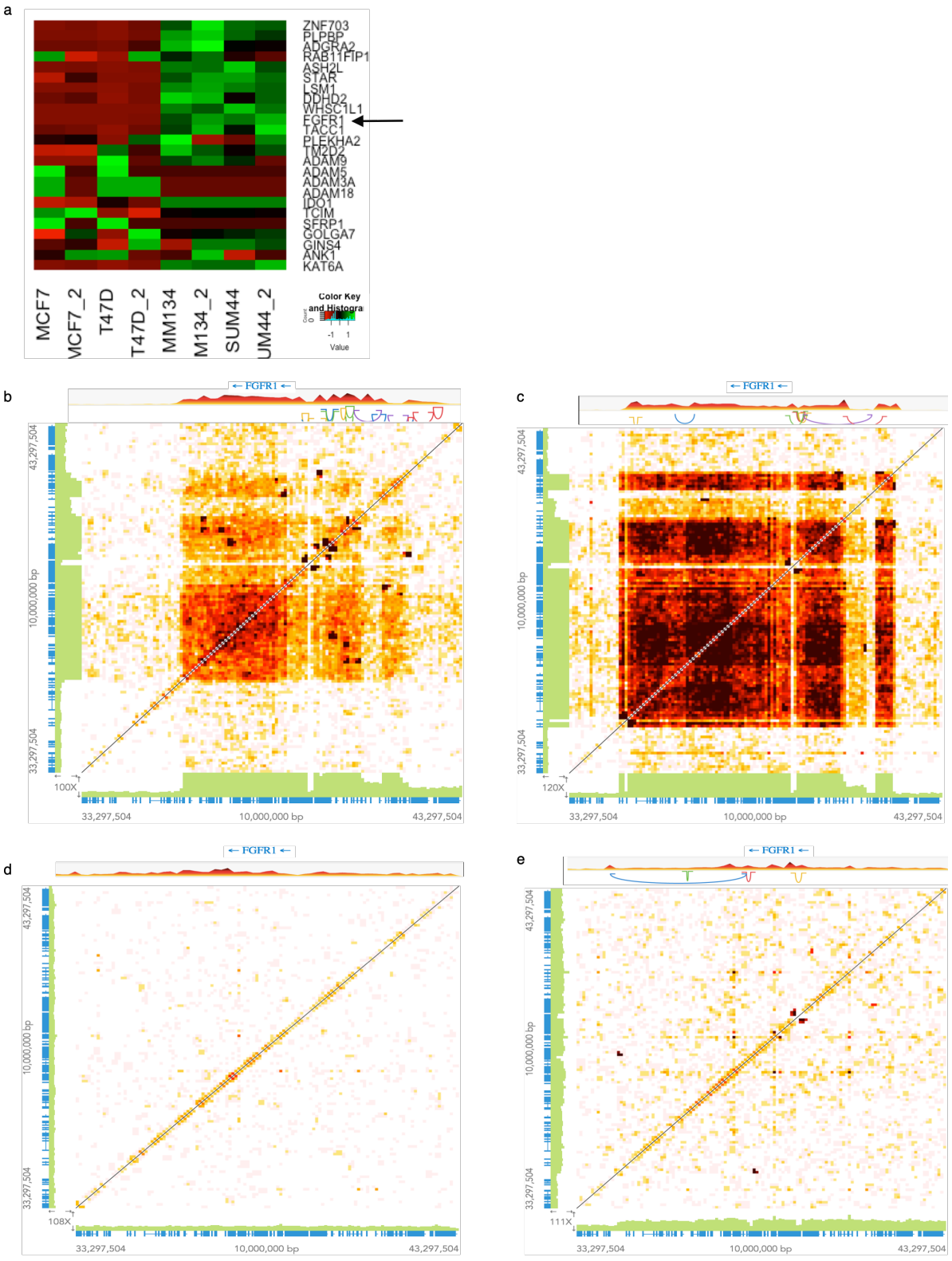


Figure 5 I FGFR1 amplification in ILC cell lines a. Heat map of the expression levels of genes mapping to the amplicon on 8p11.2-p12 in our cell lines. Colors represent relative levels of gene expression: high levels of expression (brightest green) and low levels or absence of expression (red) when compared with universal human reference RNA. b-e. Maxtrix view of 10x long read data for each cell line at chromosome 8. b. SUM44 c. MM134 d. MCF7 e. T47D. FGFR1 is labeled on the linear view of the chromosome tracks and we can see that in both our lobular cell lines (b,c) that there is a large area that is amplified around FGFR1 that directly corresponds to the dark boxes on the matrix plot. There are also many structural variants at the end of this amplicon. This amplification ismissing in our ductal cell lines.

Amplifications of Small Enhancers of MCF7

The increased resolution of structural variant calls using linked read data allows the identification of small events that may be missed by other technologies. Indeed, there are 207 amplifications in MCF7 (36% of amplifications) that are 250kb or smaller in size. We hypothesized that these small amplification may represent amplification of enhancer elements. Indeed, 81.6% (169 of 207) of these amplifications contain one or more H3K4me1 binding site. Furthermore, 70 of these amplifications (33.8%) contain one or more ESR1 binding sites, as detected by ChIPseq. Next, we compared the breakpoints of these amplifications to CTCF binding sites. CTCF is an insulator that has been show to mediate looping of enhancer elements. CTCF ChIPseq data for MCF7 from the ENCODE project was downloaded, and the distance from each amplification breakpoint to the closest CTCF binding sites were calculated. Whether this difference was closer than expected by chance was analyzed using two different methods. First, the amplification breakpoints, maintaining amplification distance, were shuffled 1000 times throughout the genome, and the distribution of distances to the closest CTCF binding site to shuffled break point was compared to the distribution of distances of the actual break points to CTCF binding sites. In all 1000 tries, the actual distribution was statistically significantly smaller than the random distribution (KS test, FDR < 0.01).

We also used the R package GenometricCorr to test the null hypothesis that CTCF binding sites and amplification break points are spatially independent (Figure 6). The low (0) p-value calculated in relative.distances.ks.p.value is in accordance with the observation that the break points and CTCF sites overlap. Relative.distances.ecdf.area.correlation is positive, so the CTCF sites and breakpoints are in general closer to the projection.test.p.value is zero, indicating either significant overlap or significant lack of overlap. Projection.test.lower.tail is FALSE,

meaning that we are in the upper tail of the distribution and there is significantly more overlap of the CTCF sites and breakpoints than we would expect if they were independent. Projection.test.obs.to.exp=2 confirms it. All three permutation tests give <0.01 meaning that the observed spatial relationships (absolute or relative distance apart) are significantly different than what is seen in the permutation distribution. From the p-values of the permutation distributions we cannot tell whether the query and reference intervals are significantly close together or significantly far apart. As the value of the scaled.absolute.min.distance.sum.lower.tail is TRUE, we know that the absolute distances between query and reference are consistent and small, and, finally, the jaccard.measure.lower.tail is FALSE, indicating an unexpectedly high overlap, as defined by the Jaccard measure.

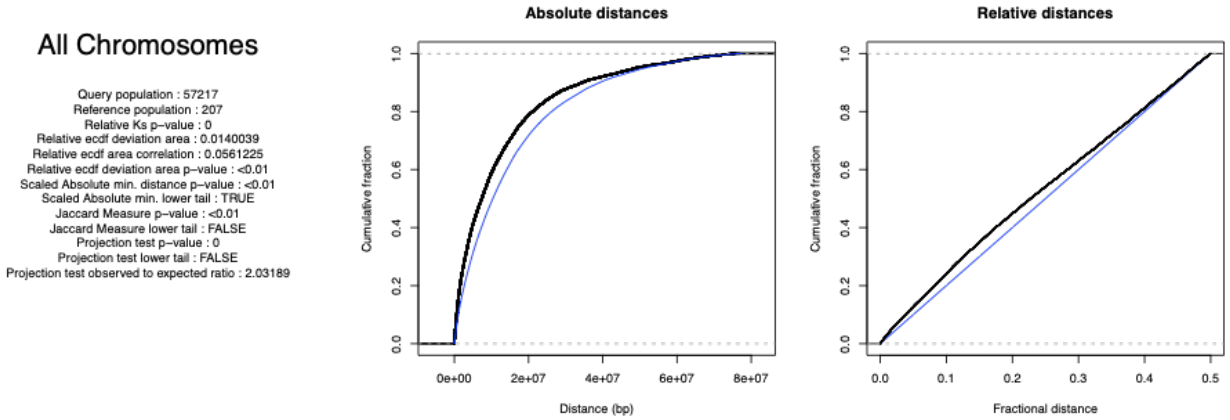


Figure 6 | Amplifications of small enhancers in MCF7: CTCF vs breakpoints. The low p-value in relative.distances.ks.p.value defines that the break points and CTCF sites overlap. Relative.distances.ecdf.area.correlation is positive, so the CTCF sites and breakpoints are in general closer. The projection.test.p.value is zero, indicating either significant overlap or significant lack of overlap. Projection.test.lower.tail is FALSE, meaning that we are in the upper tail of the distribution and there is significantly more overlap of the CTCF sites and breakpoints than we would expect if they were independent. Projection.test.obs.to.exp=2 confirms it. The value of the scaled.absolute.min.distance.sum.lower.tail is TRUE, shows that the absolute distances between query and reference are consistent and small, and, finally, the jaccard.measure.lower.tail is FALSE, indicating an unexpectedly high overlap.

CHAPTER 3

METHODS

Cell Culture

MCF7, T47D and MM134 cell were obtained from Carlos Arteaga, Vanderbilt University. MCF7 cells were grown in DMEM supplemented with 10% heat-inactivated fetal bovine serum (FBS), 50 U/mL penicillin, and 50 mg/mL streptomycin. T47D cells were grown in RPMI supplemented with 10% heat-inactivated FBS, 0.002% insulin, and 50 U/mL penicillin, and 50 mg/mL streptomycin. MM134 cells were grown in 1:1 ratio of DMEM and L-15 supplemented with 10% heat-inactivated FBS, 50 U/mL penicillin, and 50 mg/mL streptomycin. SUM44 cells were purchased from Asterand Bioscience and grown according to their instructions. SUM44 cells were grown in Hams F-12 supplemented with 2% heat-inactivated FBS, 1g/L BSA, 5mM ethanolamine, 10nM HEPES, 1ug/ml hydrocortisone, 5ug/ml insulin, 50nM sodium selenite, 5ug/ml apo-Transferrin, 10nM Triiodo Thyronine

10X Sequencing

HMW DNA preparation for 10X Genomics WGS

High molecular weight (HMW) DNA was extracted from all cell lines using the Salting Out Method (10X Genomics). Starting with 1.5 x 10^6 live cells, pellet and lyse overnight with proteinase K at 37^o. DNA is then extracted using Eppendorf DNA LoBind tubes. Extracted genomic DNA was analyzed via TapeStation at VANTAGE to check size and integrity. TapeStation gives us a DNA Integrity Number (DIN) wherein a high DIN (scale of 1 to 10) is

indicative of highly intact DNA and a low DIN of degraded DNA. All samples had a DIN >9 and a mean DNA size >50 kb.

10X Genomics WGS library construction

10X WGS Libraries were constructed at VANTAGE using the 10X Chromium protocol (10X Genomics), starting with 1.2 ng of DNA for each sample. The finished libraries were sequenced to ~30X coverage on an Illumina HiSeq3000 platform. The resulting sequencing base call files (BCLs) were processed by the Long Ranger Pipeline (10X Genomics) for alignment, structural variant discovery, and phasing.

10X Genomics WGS- Long Ranger pipeline

Samples were demultiplexed and paired end fastq files with matching barcode index files were generated with the Long Ranger (v2.2.2) mkfastq functions. The Long Ranger Pipeline (10X Genomics) was run on our four samples. This pipeline performs alignment using the Lariat aligner, which bins read-pairs containing the same molecular barcode identifier into read clouds and performs the alignment of these read-pairs simultaneously with the prior knowledge that these read-pairs originate from a small number of larger DNA molecules. The output of the pipeline included barcoded and phased BAM files, VCFs, SV VCFs, BEDPEs and a Loupe file for data visualization.

RNA-Sequencing

RNA Collection

Cells were harvested at steady-state and RNA was purified using the RNeasy kit (Qiagen). RNA samples were subjected to Turbo DNAse (Thermo Scientific) and RNA SpikeIns were added as controls.

RNA Library building and analysis

RNA samples were assessed for quality at VANTAGE core at Vanderbilt; samples with RNA integrity number of 7 or greater were used to generate RNA libraries using NEBNext Poly(A) mRNA Magnetic Isolation and NEBNext Ultra RNA Library Prep Kit for Illumina. Libraries were sequenced at VANTAGE with PE75 to a depth of approximately 30 million reads per sample on an Illumina HiSeq3000. RNA-seq reads were aligned to the human genome (hg19) with splice-aware aligner STAR and number of reads was quantified and normalized using HTSeq (Dobin et al., 2013). Differential expression analysis was performed in R using DESeq2.

ChIP-Sequencing

ChIP was done using MCF7, T47D, MM134 and SUM44 and in DMEM, RPMI, DMEM:L-15, or Hams F-12 respectively. Cells were grown to 80% confluency, washed 3 times in ice-cold PBS, and then fixed for 10 minutes at room temperature using 7% formaldehyde, followed by quenching with 2.5 mol/L glycine. Cells were first lysed using Farnham lysis buffer and then with nuclei lysis buffer (50 mmol/L Tris-HCl pH 8.0, 10 mmol/L EDTA pH 8.0, 1% SDS). Chromatin was sonicated using a Covaris LE220 with the following conditions: 35 minutes at peak power 350, duty factor 15, 200 cycles/burst, and average power 52.5; 200 mL of the chromatin was saved for input. Sonicated chromatin was diluted using ChIP Dilution Buffer (50 mmol/L Tris-HCl pH 8.0, 0.167mol/L NaCl, 1.1% Triton X-100, 0.11% sodium deoxycholate), RIPA-150, protease inhibitors, and sodium butyrate. ERα antibodies (Santa Cruz sc-543X) were linked to magnetic anti-rabbit Dynabeads (sheep anti-rabbit IgG M-280 from Life Technologies), and then incubated with chromatin for >12 hours at 4C. Immunoprecipitate (IP)

was washed with the following buffers (RIPA-150, RIPA-500, RIPALiCl, and TE Buffer) for 5 minutes each. Chromatin-IPs were eluted from the beads, treated with RNase A at 65°C with shaking for 4 hours to reverse crosslinking, followed by proteinase-K treatment at 55°C for 1 hour. Next, DNA was purified using phenol–chloroform extraction, followed by ethanol precipitation and subsequent quantification by Qubit. Standard Illumina ChIPseq Library Kits were used to build sequencing libraries. Libraries were sequenced at Vanderbilt Technologies for Advanced Genomics (VANTAGE). The fastq files were aligned to human genome version 19 by BWA (Burrows–Wheeler aligner). Peaks were called against matching input using SPP according to ENCODE best practices.

Data visualization

For visualizing the genome-wide landscape of SVs, we applied the perl Circos package v0.69-6 (Krzywinski et al., 2009). Specifically for visualization of results of structural variants and gene expression, we applied custom R code using R-packages ggplot2 (v2.2.1) and heatmap.2 (v3.0.1).

CHAPTER 4

CONCLUSION

Breast cancer is the most commonly diagnosed, non-skin cancer, in the developed world in women.[33] In fact 1 out of every 8 women will be diagnosed with breast cancer within their lifetime in the United States, accounting for over 40,000 deaths annually.[34] Breast cancers exhibit a large range of morphological features, immunohistochemial profiles, and histological subtypes that can dictate their clinical course of treatment and as well as outcome. Breast cancers can be subclassified based on histologic criteria (ductal verses lobular) as well as molecular profiling (estrogen receptor and progesterone receptor expression, HER2 amplification or triple negative). Breast cancers that are driven by ER account for 70% of all breast cancers and while anti-estrogen therapies have been successful in improving outcomes, ER+ breast cancers are very heterogeneous.[35,36] In addition to molecular profiling, breast cancer is a histological diagnosis. IDCs account for 70-80% of breast cancers while ILC account for 10-15%.[34] While these subtypes differ in histology they also have different genetic signatures. Previous work has shown different genomic landscapes between the two subtypes with different functional genomic characteristics.[37] These distinct molecular portraits between the histological subtypes of breast cancer highlight the need for individualized therapies based on histology.

Advances in long-read sequencing technologies have produced better quality reference genome assemblies and identified previously hidden genomic variation in human genomes.[38,39] Previous studies have compared long-read sequencing against short-read Illumina paired-end sequencing to investigate the performance of long and short reads for cancer genome analysis.[40]

These studies found that long-read sequencing can expose complex structural variations with more certainty that short-read sequencing. This is due to better mapping through repetitive elements that are often next to structural variants. Long-read sequencing, such as 10X phased sequencing, is a valuable resource to capture the complexity of structural variations on both the genomic and transcriptomic levels.

The work presented began as we were investigating recurrently mutated transcription factors in ER+ breast cancer. We became aware of the distinct genomic differences between histological subtypes during our analysis and wanted to investigate other structural variant differences between the two subtypes. Thus, we were able to utilize a rather new and underutilized technology to better identify complex genomic differences between histological subtypes previously unreported.

This research aimed to identify structural variation between histological subtypes of ER+ breast cancer. Using 10X genomics as well as RNA-seq we wanted to understand the complete landscape of structural variant changes of invasive ductal and invasive lobular carcinomas. Additionally we aimed to better understand ER regulation in our histological subtypes utilizing ER ChIP-seq. In the work that I presented we have identified over 2,600 large structural variants between our four cell lines using long-read sequencing. This type of sequencing has never been done before for T47D, MM134 or SUM44.

Based on whole genome sequencing and RNA-sequencing analysis it can be concluded that structural variants between cancer genomes are diverse. The results indicated some similarities between lobular cell lines such as the 8-11 translocations, which are also evident in TCGA samples. Additionally, 316 variants were identified that are shared between ductal cell lines. However for the most part, these cell lines have a large number of unique structural

variants (1,682, which is 68% of all the structural variants we identified). Therefore, we can conclude that structural variation is varied between different breast cancer cell lines.

Future studies should focus on further characterizing specific structural variants of interest. While we have identified 316 common variants between IDC and 10 common variants between ILC, the need to determine direct functional consequences of those variants are very important. Are these silent variations or are they having an impact on cancer progression. Additionally since we only have direct similarity between 10 variants in ILC could they be a defining characteristic of the histological subtype. Furthermore, the identified 8-11 translocation requires better characterization. While we identified this structure and the genes involved in our lobular cell lines as well as TCGA samples in this translocation, we have yet to determine if these contribute functionally to the lobular phenotype. Additionally, we further characterized the previously reported CCDC170-ESR1 in MCF7 cells by identifying an amplification of an enhancer region. While our data identifies this fusion in our 10X data we also looked at samples in TCGA that had the fusion and were better able to understand gene expression levels of ER in this context. I think further studies could be implemented on MCF7 cells to understand how the amplification is influencing ER functions. For example, we could cut out the amplified region using CRISPR and if it is driving ER expression, ER expression levels will drastically drop. As a control you could cut this region out of T47D, in which it is not amplified and you would see a much less reduction in ER expression levels.

It is important to consider that we only sequencing four cell lines (two of each histological subtype). Better conclusions could be drawn from larger cohorts but unfortunately many more studies are limited by the number of immortalized cell lines (especially for lobular). The next steps would be to use our technological approach and apply it on patient samples.

While we have the large database of TCGA that include 660 ER+ breast cancer tumors, these were only sequenced using short-read technologies. It would be incredibly useful to take patient tumors of both histological subtypes and use long-read sequencing to see if we get similar numbers of structural variants as well and any overlapping structural variants as our cell lines.

In summary this work has expanded our knowledge of the diversity of structural variation not only in cancer genomes but also specifically between the two histological subtypes of ER+ breast cancer. This type of analysis in invaluable in the field of cancer genomics because it sheds light on previously unreported structural variants as well as further characterizing ones that have already been identified. This analysis also shows the advancement of sequencing technologies. While short-read Illumina based sequencing is commonplace for sequencing needs, we need to take into account all the information it is missing and new technology that can fill in those gaps in knowledge.

REFERENCES

1. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA* **321**, 288–300 (2019).
2. Sharma, G. N., Dave, R., Sanadya, J., Sharma, P. & Sharma, K. K. VARIOUS TYPES AND MANAGEMENT OF BREAST CANCER: AN OVERVIEW. *J. Adv. Pharm. Technol. Res.* **1**, 109–126 (2010).
3. Nasrazadani, A., Thomas, R. A., Oesterreich, S. & Lee, A. V. Precision Medicine in Hormone Receptor-Positive Breast Cancer. *Front. Oncol.* **8**, (2018).
4. Enmark, E. & Gustafsson, J.-Å. Oestrogen receptors – an overview. *J. Intern. Med.* **246**, 133–138 (1999).
5. Johnston, S. J. & Cheung, K.-L. Endocrine Therapy for Breast Cancer: A Model of Hormonal Manipulation. *Oncol. Ther.* **6**, 141–156 (2018).
6. Turashvili, G. & Brogi, E. Tumor Heterogeneity in Breast Cancer. *Front. Med.* **4**, (2017).
7. Makki, J. Diversity of Breast Carcinoma: Histological Subtypes and Clinical Relevance. *Clin. Med. Insights Pathol.* **8**, 23–31 (2015).
8. Lobular breast carcinoma and its variants- ClinicalKey. Available at: https://www.clinicalkey.com/#!/content/playContent/1-s2.0-S0740257009000999?returnurl=https:%2F%2Flinkinghub.elsevier.com%2Fretrieve%2Fpii%2FS0740257009000999%3Fshowall%3Dtrue&referrer=https:%2F%2Fwww.ncbi.nlm.nih.gov%2F. (Accessed: 8th August 2019)
9. Barroso-Sousa, R. & Metzger-Filho, O. Differences between invasive lobular and invasive ductal carcinoma of the breast: results and therapeutic implications. *Ther. Adv. Med. Oncol.* **8**, 261–266 (2016).
10. Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506–519 (2015).
11. Hanahan, D. & Weinberg, R. A. Hallmarks of Cancer: The Next Generation. *Cell* **144**, 646–674 (2011).
12. Zhang, Y. *et al.* Copy Number Alterations that Predict Metastatic Capability of Human Breast Cancer. *Cancer Res.* **69**, 3795–3801 (2009).
13. Chi, C., Murphy, L. C. & Hu, P. Recurrent copy number alterations in young women with breast cancer. *Oncotarget* **9**, 11541–11558 (2018).
14. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups | Nature. Available at: https://www.nature.com/articles/nature10983. (Accessed: 2nd August 2019)
15. Zahir, F. R. & Marra, M. A. Use of Affymetrix Arrays in the Diagnosis of Gene Copy-Number Variation. *Curr. Protoc. Hum. Genet.* **85**, 8.13.1-8.13.13 (2015).
16. Elyanow, R., Wu, H.-T. & Raphael, B. J. *Identifying structural variants using linked-read sequencing data*. (Genomics, 2017). doi:10.1101/190454
17. Zhang, L., Feizi, N., Chi, C. & Hu, P. Association Analysis of Somatic Copy Number Alteration Burden With Breast Cancer Survival. *Front. Genet.* **9**, (2018).
18. Marks, P. *et al.* Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res.* **29**, 635–645 (2019).
19. *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*. (Elsevier, 2018).

20. Zheng, G. X. Y. *et al.* Haplotyping germline and cancer genomes using high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311 (2016).
21. Raghupathy, N. *et al.* Hierarchical analysis of RNA-seq reads improves the accuracy of allele-specific expression. *Bioinformatics* **34**, 2177–2184 (2018).
22. Sikora, M. J. *et al.* Invasive lobular carcinoma cell lines are characterized by unique estrogen-mediated gene expression patterns and altered tamoxifen response. *Cancer Res.* **74**, 1463–1474 (2014).
23. Lei, J. T., Gou, X. & Ellis, M. J. ESR1 fusions drive endocrine therapy resistance and metastasis in breast cancer. *Mol. Cell. Oncol.* **5**, e1526005 (2018).
24. ESR1 alterations and metastasis in estrogen receptor positive breast cancer. Available at: https://jcmtjournal.com/article/view/3064.
25. Veeraraghavan, J. *et al.* Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of estrogen-receptor positive breast cancers. *Nat. Commun.* **5**, 4577 (2014).
26. Giltnane, J. M. *et al.* Genomic profiling of ER+ breast cancers after short-term estrogen suppression reveals alterations associated with endocrine resistance. *Sci. Transl. Med.* **9**, (2017).
27. Nambiar, M., Kari, V. & Raghavan, S. C. Chromosomal translocations in cancer. *Biochim. Biophys. Acta BBA - Rev. Cancer* **1786**, 139–152 (2008).
28. Rowley, J. D. A New Consistent Chromosomal Abnormality in Chronic Myelogenous Leukaemia identified by Quinacrine Fluorescence and Giemsa Staining. *Nature* **243**, 290–293 (1973).
29. Perez-Garcia, J., Muñoz-Couselo, E., Soberino, J., Racca, F. & Cortes, J. Targeting FGFR pathway in breast cancer. *The Breast* **37**, 126–133 (2018).
30. Piasecka, D. *et al.* FGFs/FGFRs-dependent signalling in regulation of steroid hormone receptors – implications for therapy of luminal breast cancer. *J. Exp. Clin. Cancer Res.* **38**, 230 (2019).
31. Turner, N. *et al.* FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer. *Cancer Res.* **70**, 2085–2094 (2010).
32. Sikora, M. J. *et al.* Invasive lobular carcinoma cell lines are characterized by unique estrogen-mediated gene expression patterns and altered tamoxifen response. *Cancer Res.* **74**, 1463–1474 (2014).
33. Waks, A. G. & Winer, E. P. Breast Cancer Treatment: A Review. *JAMA* **321**, 288–300 (2019).
34. Sharma, G. N., Dave, R., Sanadya, J., Sharma, P. & Sharma, K. K. VARIOUS TYPES AND MANAGEMENT OF BREAST CANCER: AN OVERVIEW. *J. Adv. Pharm. Technol. Res.* **1**, 109–126 (2010).
35. Nasrazadani, A., Thomas, R. A., Oesterreich, S. & Lee, A. V. Precision Medicine in Hormone Receptor-Positive Breast Cancer. *Front. Oncol.* **8**, (2018).
36. Johnston, S. J. & Cheung, K.-L. Endocrine Therapy for Breast Cancer: A Model of Hormonal Manipulation. *Oncol. Ther.* **6**, 141–156 (2018).
37. Ciriello, G. *et al.* Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell* **163**, 506–519 (2015).
38. Chaisson, M. J. P. *et al.* Resolving the complexity of the human genome using single-molecule sequencing. *Nature* **517**, 608–611 (2015).
39. Pendleton, M. *et al.* Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat. Methods* **12**, 780–786 (2015).

40. Nattestad, M. *et al.* Complex rearrangements and oncogene amplifications revealed by long-read DNA and RNA sequencing of a breast cancer cell line. *Genome Res.* (2018). doi:10.1101/gr.231100.117
41. C Desmedt, G Zoppoli, G Gundem, *et al.*Genomic characterization of primary invasive lobular breast cancer. J Clin Oncol, 34 (2016), pp. 1872-1881
42. Cowin, P. A. *et al.* LRP1B deletion in high-grade serous ovarian cancers is associated with acquired chemotherapy resistance to liposomal doxorubicin. *Cancer Res.* **72**, 4060–4073 (2012)
43. Dean L. Fluorouracil Therapy and DPYD Genotype. 2016 Nov 3. In: Pratt V, McLeod H, Rubinstein W, et al., editors. Medical Genetics Summaries [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2012-.
44. Sethuraman, A., Brown, M., Krutilina, R., Wu, Z. H., Seagroves, T. N., Pfeffer, L. M., & Fan, M. (2018). BHLHE40 confers a pro‐survival and pro‐metastatic phenotype to breast cancer cells by modulating HBEGF secretion. *Breast Cancer Research*, **20**(1), 117.
45. Harding, H. et al. An integrated stress response regulates amino acid metabolism and resistance to oxidative stress. *Mol. Cell* **11**, 619–633 (2003).
46. Cartwright, T., Perkins, N. D., Wilson, C., NFKB1: a suppressor of inflammation, ageing and cancer. *FEBS J*. 2015
47. Takeda DY, et al. A somatically acquired enhancer of the androgen receptor is a noncoding driver in advanced prostate cancer. *Cell*. 2018;174(2):422–432.e13.