Sparse canonical correlation to identify breast cancer related genes regulated by copy number aberrations

Diptavo Dutta[1], Ananda Sen[2,3], Jaya Satagopan[4]

1. Department of Biostatistics, Johns Hopkins University
2. Department of Biostatistics, University of Michigan
3. Department of Family Medicine, University of Michigan
4. Department of Biostatistics and Epidemiology, Rutgers University

Abstract

Background: Copy number aberrations (CNA) have proved to be of clinical and therapeutic significance for many diseases including breast cancer, since they drive numerous key underlying biological processes, by regulating molecular phenotypes like gene expression and others. To comprehensively assess the effect of CNAs, it is not sufficient to only identify significant CNA-gene expression pairs, but also to identify the overall gene networks and regulatory structures that are influenced by CNAs, subsequently producing change in outcomes.

Methods: In this article, we adopt a two-step analysis approach to identify CNA regulated genes whose expression levels affect breast cancer related outcomes: (1) we identify gene modules that are regulated by CNAs through sparse canonical correlation analysis (sCCA) which selects a set of closely located CNAs that regulates the expression levels of selected genes. (2) then, we use a using generalized linear model, to identify which genes within the gene modules are associated with breast cancer related outcomes.

Results: Analyzing clinical and genomic data on 1904 breast cancer patients from the METABRIC study, we found 14 gene modules to be regulated by groups of proximally located CNA sites. The identification of gene modules was further validated using independent data on individuals in a study of breast invasive carcinoma from The Cancer Genome Atlas (TCGA). Association analysis on 7 different breast cancer related outcomes identified several novel and interpretable regulatory associations which highlights how CNA can impact key biological pathways and process in context of breast cancer. Through downstream analysis of two example outcomes: estrogen receptor status and overall survival, we show that the identified genes were enriched in relevant biological pathways and the key advantage of our method is that we additionally identify the CNA that regulate these genes. Due to the availability of multiple types of outcomes, we further meta-analyzed the results to identify genes that had potentially associations with multiple outcomes.

Conclusions: Overall we present a generalizable analysis approach to identify genes associated to different outcomes that are regulated by sets of CNA and can further be used to combine results across various types of outcomes. The results show that our method can identify novel and interpretable associations, by providing mechanistic insights on how the effects of CNA are cascaded via gene expression to impact breast cancer and related outcomes.

Introduction

With rapid advancement in sequencing technologies, large-scale genomic studies have identified somatic variations that influence the risk of breast cancer (BrC) [1–4]. These findings have been critical in enhancing our understanding of the underlying biology of breast tumorigenesis[5]. This has further led to the development of diagnostic, prognostic and screening tests and more importantly, the development of effective targeted therapies and innovative prevention strategies[6,7].

Among the genetic variations that have been studied in context to breast cancer and related outcomes, copy number aberrations (CNAs), such as copy number gains and losses, constitute an important and widely studied class[8]. Several studies have shown that cancer genomes have an enriched burden of copy number aberrations[9–11]. Additionally, such alterations have been shown to harbor or be proximal to important oncogenes and tumor suppressor genes[12]. Thus, CNAs might directly regulate cellular growth pathways and other gene sets that impact key biological mechanisms that influence cancer-related outcomes. In fact, studies associating CNAs with breast cancer have considerably advanced our knowledge of breast cancer biology[13,14], with translational efforts leading to advances in the clinic. Identification of driver CNAs and understanding how they affect gene expression, eventually impacting breast cancer and related outcomes, however, remains a challenge.

It is popularly hypothesized that somatic variations like CNA, regulate a complex network of intermediate molecular phenotypes, like gene expressions. Identifying such genetic regulatory structures can inform downstream consequences of CNA and the broad biological functions and mechanisms that are affected[15,16]. However, standard association mapping methods typically aim to identify only significant pairs of CNA-gene expression, which does not provide insight into the overall gene modules and pathways that might be influenced. Thus, to achieve a comprehensive map of the regulatory networks, via which the effects of CNA might be cascaded, it is not sufficient to identify only individual CNA-gene pairs that are associated. Mapping CNAs or groups of closely related CNAs that impact gene networks or modules has significant potential for providing practical insights into the regulatory impact of CNA on gene networks, which can be further examined in translational investigations for devising novel therapeutic strategies.

In this article, we present a two-step analysis framework that aims to map sets of CNA that regulate gene expression to affect breast cancer related outcomes: (1) In the first step, we identify gene modules that are regulated by CNAs by employing sparse canonical correlation analysis (sCCA)[17,18] which selects a group of closely located CNAs that regulates the expression levels of selected genes.

This is an unsupervised step in the sense that it is agnostic of the breast cancer related outcomes and identifies overall patterns of gene-expression regulation by CNA. (2) Given the gene modules identified in the sCCA analysis, in the next step we use a standard association test using generalized linear model, to identify which genes within a particular module are associated with breast cancer related outcomes.

This approach is particularly amenable to interpretation since not only it identifies the genes whose expression levels impacts breast cancer related outcomes but also identifies the set of CNA which potentially regulates them. Broadly, this approach can provide a mechanistic insight on which gene networks and biological processes are regulated by CNA to influence breast cancer related outcomes. We analyze data on 1904 breast cancer patient whose CNA and gene-expression profiling was performed as a part of the METABRIC study[8,19] (See Supplementary Methods for details on the study). We identified 14 gene modules regulated by groups of CNA across the genome which included several trans-associations as well. We further validated the identification of gene modules in an independent data on individuals in a study of breast invasive carcinoma from The Cancer Genome Atlas (TCGA), obtained from the cBioPortal catalog[20]. Subsequent association analysis on 7 different BrC related outcomes showed that novel and interpretable regulatory associations were identified which highlights how CNA disrupt and influence important biological functions and process in context of BrC. We have demonstrated the utility of our approach using examples of two outcomes: estrogen receptor status and overall survival. Due to the availability of multiple types of outcomes, we further meta-analyzed the results to identify genes that had potentially associations with multiple outcomes.

Overview of methods

To describe our approach, we assume that we have individual level data for n individuals on p copy number aberrations (CNA) and q gene-expressions.

Step 1. Identifying Gene modules through sCCA

We first aim to identify gene modules regulated by CNA, by mapping groups of CNA to groups of associated gene expressions using sparse canonical correlation analysis (sCCA). sCCA identifies approximate orthogonal gene modules that are regulated by CNA. This step is agnostic of any phenotypic information or outcomes. For *n* individuals, let $\boldsymbol{G}^{n\times p}$ be the matrix for *p* sites of copy number aberration (CNA) sites with $\boldsymbol{G}_{ij}$ being the number of insertion or deletions for individual i at site j, and $\boldsymbol{E}^{n\times q}$ be the normalized gene-expression levels for *q* genes across n individuals. Sparse

canonical correlation analysis (sCCA) identifies sparse linear combinations of CNA ($\boldsymbol{u}^{\boldsymbol{p}\times\boldsymbol{1}}$; termed CNA component) and gene-expressions ($\boldsymbol{v}^{\boldsymbol{g}\times\boldsymbol{1}}$; termed gene component) such that the correlation between $\boldsymbol{Gu}$ and $\boldsymbol{Ev}$ is maximized i.e.,

$$(u, v) = argmax\ \tilde{v}^T \tilde{E}^T \tilde{G} \tilde{u}$$

$$\text{under } ||\tilde{u}||_1 \leq c_u \ ; ||\tilde{v}||_1 \leq c_v \text{ and } ||\tilde{u}||_2 = 1, ||\tilde{v}||_2 = 1$$

where $||.||_h$ denotes the L_h norm and $\widetilde{\boldsymbol{G}}$ (or $\widetilde{\boldsymbol{E}}$)denotes the normalized version of the corresponding matrix. The subsequent pairs of sCCA components are obtained similarly by matrix deflation and under the constraint of being uncorrelated or orthogonal to the previous components. Ideally each pair of sCCA components selects a sparse set of CNA sites that regulate the expression of a sparse set of genes across the genome, denoted by the non-zero elements in u and v respectively. Overall, the sCCA aggregates multiple associations between the selected CNA and genes and hence represents principal regulation or association patterns. Additionally, due to the orthogonality constraint each pair of sCCA component reflect approximately an independent or orthogonal pattern of regulation. c_u (and c_v) represent the sparsity parameters for the CNA and gene components respectively. To facilitate interpretation, we choose the sparsity parameters such that there is no overlap between the CNA selected in the components. (See Supplementary Methods for details).

Step 2. Association with outcomes

Given the gene-modules identified in Step 1, we now identify which genes within these modules are associated with the outcomes of interest.

Univariate outcomes. Let $\tilde{E}_k$ be the $n \times r_k$ matrix of normalized gene-expressions for the genes selected in sCCA component *k*, where $r_k = ||v_k||_0$ and v_k denotes the gene-component of the kth sCCA component. We use the following generalized linear model to associate the r_k genes to a phenotype y as

$$g[E(y)] = \beta_0 + \tilde{E}_k \beta$$

Where $g[.]$ is a canonical link function and β, β_0 are regression parameters. For each of the gene modules identified by sCCA, we perform the association analysis and record gene-specific p-values and obtain the false discovery rates (FDR). Genes with with FDR < 0.05 are declared to be significantly associated with the outcome.

Multivariate outcomes. If multiple, potentially correlated, outcomes are available for the individuals, we can meta-analyze results across the multiple outcomes to identify genes that are possibly associated to more than one outcome. Let $p_1, p_2, \ldots, p_s$ be the univariate p-values for a particular gene for s outcomes, from the previous univariate association analysis. These p-values are likely to be correlated due to potential correlation between the outcomes. We perform a cauchy-transformed meta-analysis[21] which has been shown to maintain correct false positive rate in presence of correlation as well[22,23]. We transform each of the p-values to a cauchy variable as

$$c_i = \tan(\pi(p_i - 0.5))$$

The test statistic is the unweighted mean of these transformed variables which follows a standard cauchy distribution, under the null hypothesis of no association, irrespective of the correlation between the outcomes[21].

$$T = \frac{1}{s}\sum_{i=1}^{s} c_i \sim Cauchy(0,1)$$

The overall p-value can be calculated by inverting the cumulative density function of the standard cauchy distribution.

Results

We started with 1,904 individuals who had complete data at 22,544 CNA sites and expression level data for 24,360 genes. Sparse canonical correlation analysis (See Methods) identified 14 gene modules through the sCCA components. For the purpose of this article, we will use the terms *modules* and *networks* synonymously to denote the collection of genes selected in a gene component and *set* to denote the CNA sites selected in a CNA component. Across the 14 gene modules, sCCA selects 831 genes, whose expression levels are regulated by 1,976 CNA sites overall (Table 1). In general, for each sCCA component, the CNA component selects CNA sites located in a small sub-region within a chromosome (Figure 1A). Our sCCA analysis was agnostic of the physical location of the CNA in the genome and hence the sCCA algorithm is not guided or biased towards selecting positionally proximal CNA. However, due to the high correlation between nearby CNA, each CNA component selects a smaller subregion in chromosome of high correlated CNA which might have regulatory effects on the gene selected in the corresponding gene component. For example, the 115 CNA sites selected in CNA component 4, were located on chromosome 17q11.2-q21.32 region. The corresponding gene components can then be viewed as the gene module having strong association, on an average, with the selected CNA sites and mediates their effects. In other words, the 67 genes selected in gene component 4 would be the gene module (or network) that is regulated by the 115 copy number aberrations selected in the CNA component 4. In general, we expect the regulatory

structures captured by the sCCA components to be approximately independent. However, we notice that the expression levels of genes selected in gene component 8 has a higher correlation with the CNA selected in CNA component 2 (Figure 1B). It is to be noted that CNA components 2 & 8 defined highly proximal regions in chromosome 8. Hence, the correlation between gene module 8 & CNA set 2 is not unexpected due to LD and/or possible long range regulatory activity. This indicates that genes modules 2 and 8 are possibly coregulated by the CNA selected in the corresponding CNA components. Overall, a CNA components defines a chromosomal subregion which has potentially multiple independent regulatory effects on the gene module identified by the corresponding gene component. The advantage of the sCCA in this application is that it can aggregate multiple, possibly weaker association to select groups of CNA associated with genes modules (See Supplementary Table 1-2 for full list of CNA and genes selected).

Gene modules capture cis and trans effects. Through the identification of gene modules, we capture regulatory effects of CNA. In general, we found that most of the associations aggregated by the sCCA components identified effects of CNA sites on nearby (cis) gene expression. On average, 44.8% of the genes selected in each sCCA component also has a CNA in or near the same gene selected in the respective CNA component. This is expected since cis effects are known to be much stronger compared to distal (trans) effects and would have a direct regulatory effect on the expression level of a nearby gene. However, several examples of distal (trans) regulatory effects on expressions of genes on different chromosome were also identified in the gene modules (Figure 2A-B). On average, 3.2% of the genes selected in the gene components were on a different chromosome than the corresponding CNA component. Further, on average 15.9% of the genes selected in the gene components were more that 10Mb away from the sub-region of chromosome selected by the corresponding CNA component, indicating long range regulatory effects (Table 1). For example, among the 67 genes selected in component 4, 9 genes are on different chromosomes and an additional 4 genes are outside of the region 17q11.2-q21.32, which contains the CNA selected by CNA component 4. We found possible mechanistic explanations for several such distal associations in existing genomic and profiling data. For example, gene component 4, selects atlastin GTPase 3 gene (*ATL3*) on chromosome 11. *ATL3* is a downstream target for transcription factor Signal Transducer and Activator of Transcription 3 (*STAT3*) in ENCODE transcription factor database[24,25]. Interestingly, a copy number aberration of *STAT3* was selected among the CNA sites in CNA component 4, which suggests a possible cis-mediation mechanism for the association of this and other nearby CNA sites with *ATL3*.

Evidence of coregulation. To further validate whether the genes selected by the 14 significant sCCA components had any overall evidence of biological coregulation as well, we used large-scale transcription factor databases from the ENCODE study[24] and existing ChIP-chip, ChIP-seq, and several other transcription factor binding site profiling experiments (ChEA)[26,27]. For the 181 transcription factors and their downstream targets reported in ENCODE, we found that, across the 14 gene module identified though sCCA components, on an average 67.3% of the genes were downstream targets for more than 20 transcription factors. For ChEA, which reports data on 202 TFs and their downstream targets, we found similarly that on average 65.1% genes were downstream targets for more than 20 transcription factors. This provides implicit evidence that a large proportion of the genes selected by the sCCA components might have evidence of being coregulated by TFs and the identification of gene modules using sCCA analysis can successfully detect such patterns of coregulation as an independent line of evidence.

Replication of Gene Modules using TCGA breast invasive carcinoma data. Selection of genes and CNA can be influenced and biased if there are systematic biases and batch effects. So, we investigated whether the gene modules and CNA sites identified through sCCA, were replicable in an external dataset. For that, we used the TCGA breast invasive carcinoma data (See Supplementary Methods for details on the study), which reports data on CNA sites and gene expression in primary breast tumor tissue for more than 1,000 breast cancer patients. We adopted a resampling-based procedure to test whether the sCCA components represented gene modules and CNA sets that had stronger association than expected at random. For a given gene module (selected through a gene component), we evaluated whether the observed average squared correlation between these genes and CNA selected in corresponding CNA component were higher than what is expected at random. Similarly, for a set of CNA (selected through a CNA component), we evaluated whether the observed average squared correlation between these CNA and genes selected in corresponding gene component were higher than what is expected at random. We found that among the gene modules and CNA sets selected in METABRIC and present in TCGA, the average correlation for all the 14 components were significantly (p-value < 0.05) higher than expected (Supplementary Figure S1). Further 10 of these components were strongly significant as well (p-value < 0.001). Such a result is not unexpected as the sCCA components include a majority of cis effects. Further, this also suggests that the sCCA components in METABRIC possibly captured true effects replicable across different datasets and not potential artefacts and batch effects within METABRIC. (See Supplementary Methods)

Association with breast cancer related outcomes. Given the 14 gene modules obtained through sCCA analysis, we investigated whether these gene modules were associated with 7 different types of breast cancer outcomes (Table 2). At a lenient cutoff of FDR < 0.05, we found that 539 genes across the 14 modules were associated with at least one of the outcomes (Supplementary Table 3). Further, at a stringent exome-wide cutoff of p-value < 2.5×10^{-06}, we found 94 genes associated with at least 1 outcome. Subsequently, through several downstream analysis we investigated whether the genes that are significant for a given outcome indeed had external evidence of association to BrC related outcomes. Here we demonstrate the results for two distinct types of outcomes:

Estrogen Receptor (ER). Of the 1,904 individuals in the sample, 1,459 (76.6%) individuals had ER positive status. We performed logistic regression-based association tests of the 14 significant gene modules. Across the components we found that 210 genes were significant at an FDR < 0.05 and 36 genes were significant with p-value < 2.5×10^{-06}. Among the genes significantly associated with ER status, we identified known breast cancer related genes such as Microtubule Associated Protein Tau (*MAPT*), whose expression is highly associated with low sensitivity to taxanes that are important drugs for breast cancer treatment[28], and Macrophage migration inhibitory factor (*MIF*), a pro-inflammatory cytokine whose blockade reduces the aggressiveness of invasive breast cancer[29]. The advantage of our approach is that the sCCA-based model provides an explanation of the intermediate biological mechanisms. For example, among genes selected in gene component 4, we found that Dynein Axonemal Light Intermediate Chain 1 (*DNALI1*), on chromosome is associated with ER status (p-value = 4.8×10^{-04}), being trans-regulated by CNA sites on chromosome 17 selected in CNA component 4 (Figure 3A). *DNALI1* is a downstream target for transcription factors *STAT3* and *UBTF*, both of which are selected in CNA component 4. Further, there is evidence of physical interactions between the proteins resulting from *DNALI1* and *UBTF* in large protein interactions databases as well[30]. This indicates the possibility that DNALI1 mediates the effects of the CNA sites in chromosome 5 selected by CNA component 1, on ER status. Thus, not only we identify the genes whose expression levels are associated with breast cancer outcomes, we also additionally identify which CNA potentially regulate such genes.

Through pathway enrichment analysis (Table 3), we found that the genes significantly associated to ER status at FDR < 0.05, were enriched for hallmark pathways[31] like early response to estrogen, DNA repair and *MYC* targets, *MYC* being a well-known oncogene[32]. Further, in pathways curated from chemical and genetic perturbation experiments, we found that the genes were enriched for genes highly positively co-expressed with BRCA1 and BRCA2, two genes well reported to be involved in BrC[33–35]. Further, the genes were also enriched for targets of several transcription factors,

like *NELFE*[36,37], *E2F4*[38,39] and *CREB1*[40,41] which are known to play key roles in several cancers, including BrC. However, overlap of the identified significant genes with key cancer related pathways suggest a possible mechanistic explanation for the outcome. For example, of the 210 significant genes, 7 genes (*ADCY9, ABAT, MAPT, SLC9A3R1, CANT1, BCL2, FAM102A*) are in the early estrogen response hallmark pathway[31]. These 7 genes are identified as part of gene modules 4, 11, 14, 7 and 10. This indicates that the CNA selected in the corresponding CNA components, which regulates these gene components respectively, as shown in sCCA analysis, significantly changes estrogen response and can possibly be causal for ER status. This interpretability is a key advantage of our analysis approach.

Overall survival (OS). Of the 1,904 individuals in the sample, 1,109 (76.6%) individuals died during the study, with median survival time being 154 months approximately. In a cox proportional hazard model, we found 73 genes to be significant across the 14 components. Notably, several interesting distally regulated genes are identified in our analysis. For example, in sCCA component 11, we found that the expression of *CD2BP2* gene on chromosome 5 is associated with the overall survival. This gene is differentially regulated in T47D cells of BrC patients in response to tamoxifen[42], a widely used hormonal therapy drug for BrC[43]. The transcription start site for *CD2BP2* is over 27 Mb downstream from the subregion of chromosome 5 selected through the CNA component 11. The corresponding selected set of CNA in component 11 contains a TF *ZNF263*, which has a long-range regulatory effect on *CD2BP2* on the same chromosome[24] and indicates that possibly *CD2BP2* mediates the effects of the selected CNA producing significant change in overall survival probability (Figure 3B-C).

A comprehensive pathway enrichment analysis (Table 4) reveals that the selected genes are enriched in gene-sets and pathways defined by several breast cancer related perturbation experiments. For example, we found a significant enrichment of the genes associated to OS, in the genes related to adipogenesis. Enrichment was found among genes up regulated in early primary breast tumors expressing ESR1 vs the ESR1 negative ones. In addition, the genes significantly associated to OS were enriched for targets of several key cancer TFs like *ELK1*, *YY1* and *RUNX3*[44–46]. As highlighted above, through the sCCA components and the subsequent cox PH association model, we not only identify which genes are associated with OS but also detect the CNA sites regulating these gene expressions and, hence affecting OS.

Multiple outcomes. We further meta-analyzed results across all the seven BrC related outcomes to identify genes that are possibly associated to multiple outcomes. Since the outcomes are correlated and as a result the association p-values across the outcomes for each gene are correlated, it is

difficult to use standard meta-analysis for this. In fact, the effect size estimates for association models pertaining to different types of outcomes (binary and survival), would complicate the interpretation of effect size based meta-analysis. Here, we used the cauchy combination test to meta-analyze results across the seven outcomes. 72 genes were identified to be significant at the exome-wide p-value threshold of 2.5×10^{-06} (Figure 3D). At FDR < 0.05, we found 495 genes to be significantly associated to the set of 7 outcomes. Although majority of these associations were driven by significant associations with one outcome and weaker association with several others, 4 genes were also identified which had no significant association (FDR > 0.05) to any single outcome but had possibly weaker association with multiple outcomes. For example, we found *ZC3H3* gene to be significant through the cauchy combination approach which has nominal associations with the grade of tumor (p-value = 0.020) and age at diagnosis (p-value = 0.031). *PAK2*, a gene well reported to be associated to different cancers including BrC[47,48] and a target for *MYC* oncogene, had multiple weaker associations with age at diagnosis (p-value = 0.032) and overall survival (p-value = 0.005). We conducted pathway enrichment analyses with the 495 genes that were found significant at FDR < 0.05 in meta-analyses. Similarly, as before, the results show that the numerous relevant pathways and gene-sets related to BrC are significantly enriched for these genes.

Discussions

Extensive research has established that CNA are indeed important for several cancer types and subtypes, especially in breast cancers[19]. However, the intermediate mechanisms and processes via which CNA impact breast cancer related outcomes have not been conclusively established and warrants further research. In this article we have outlined a novel and generalizable analytic approach to identify how CNA regulate expression levels of gene modules that ultimately influence several breast cancer related outcomes. Our approach involves two steps: using sparse canonical correlation analysis to identify gene modules associated with sets of CNA, followed by testing association between the gene modules and cancer related outcomes. We further carried out a meta-analysis across different types of outcomes to identify genes with multiple associations. Extensive downstream analysis shows that the genes identified through our analysis have key relevance for breast and other cancers that have also been noted in other studies. Unlike these other studies, our approach also identifies CNA sets that possibly regulate the genes which in turn bring about changes in outcomes related to breast cancer.

The identification of gene modules using sCCA is a key advancement that we propose over existing work on this topic. Several authors have hypothesized that the effects of genetic variants like CNA are cascaded through complex intermediate gene network to bring about phenotypic change[15]. However, associations analysis using single CNA-gene expression pair fails to provide such a mapping to identify potential co-regulation of gene modules. Through our joint analysis approach in sCCA, we map groups of CNA to gene modules, which elucidates the concept of groupwise mapping rather than individual associations and hence is more amenable to interpretation in genetic contexts. Existing transcription factor databases and profiles show that the gene modules thus detected can potentially be coregulated which indicates that such groupwise mapping approach can identify biological regulation as well.

One of the key interpretations of the gene modules is that they represent approximately independent regulatory patterns due to the orthogonality condition imposed by sCCA. Thus, in principle, the first step in our analysis, identifies key distinct biological regulatory processes that are activated within primary breast tumor tissue. sCCA identifies the gene modules that have multiple independent regulatory associations with the corresponding CNA set. This can possibly be powerful in comparison to identifying single CNA-gene pairs, since sCCA can aggregate multiple possibly weaker association within the identified components. In fact, contrasting the gene modules identified here with those identified in normal breast or mammary tissues might provide further insight as to the variation in biological mechanisms caused by tumorigenesis. There will possibly be a plethora of regulatory associations beyond the ones identified through sCCA, which can be identified in a larger sample.

The subsequent association analysis using gene modules is relatively standard and has been used commonly. However, interpreting the association p-values needs caution since the association model is a joint regression. In general, single gene vs outcome tests are different from this since the joint model additionally adjusts for correlation between the genes and reports the p-value conditional on the gene module. This joint regression framework, thus, provides an intuitive “fine-mapping” among the genes in module with respect to an outcome, in that it identifies the genes that are associated while adjusting for the correlation within the module. While association between and outcome and a single gene can arise either due to true causality or due to correlation of the tested gene with the true causal gene, our approach accounts for the dependency and hence significant association can potentially be causal.

The multiple outcome meta-analysis[21] demonstrates another advantage of our approach. In general, complex diseases like cancers, and in particular breast cancer, can have numerous biomarkers of

disparate types. Combining results across the biomarkers can highlight overall important genes and genes affecting multiple biomarkers. However, meta-analyzing association results across them is not straightforward since the results are correlated. Further for different types of biomarkers (continuous, discrete, binary, survival etc.) the effect sizes have disparate interpretation and hence standard meta-analysis might not be appropriate. However, our cauchy combination approach alleviates these problems since it is based on p-values and not on the effect size. Further, due to the correlation agnostic property of cauchy distributions, meta-analysis of correlated p-values controls false positive rates. In principle, using this approach, meta-analysis of different types of biomarkers and outcomes are possible, if the univariate association mean model is correctly specified.

sCCA has previously been previously used for identifying patterns of regulation of genetic variants and gene expressions and other intermediate phenotypes[18,49,50]. It provides an intuitive approach to map sets of genetic variants to sets or networks of molecular phenotypes like gene expressions, protein levels, metabolites and others. In fact, this broad concept can further be extended to multi-view data sets and can be useful in integrative analysis of multi-omics data. Further research is merited in this respect.

Together, our analysis provides a comprehensive picture as to how CNA can impact different breast cancer outcomes via regulation of intermediate gene networks. If a particular gene is significantly associated with a breast cancer related outcome, we can identify which set of CNA of which genomic subregion regulates it using the identified gene modules. Further, overlap of the significant genes with several breast cancer related pathways identifies the genes within the predefined biological process is differentially regulated by CNA to bring about phenotypic change.

The analysis approach that we adopted here is highly generalizable as an overall intuitive framework. The first step pertains to groupwise mapping and identification of modules followed the association analysis in the second step. Although we adopted an sCCA approach here for its ease of interpretation, several other methods mapping sets of genetic variants to gene modules can be used instead. For example, methods for biclustering and matrix factorization can be adapted in the first step to identify gene modules. In fact, groupings based on functional annotation can also be incorporated to further strengthen the mechanistic interpretation of the identified modules. The subsequent association analysis can also be customized to address scientific questions of interest. However, one of the major advantages of our pipeline is that, in principle, the two steps can be performed on separate datasets as well. For example, in current studies large scale genomic and transcriptomic data for many individuals are available while detailed information on phenotypes and

traits might not be available. Thus, the identification of gene modules can be performed using such data while the subsequent association analysis can be carried out in a separate data.

Our analysis approach currently has several limitations. The optimal number of components in sCCA is chosen heuristically by maximizing the iterative ratio of canonical correlations rather than using any significance or enrichment tests. In the current set up, formulating analytical tests of significance is difficult and methods based on sCCA has mostly resorted to resampling methods. In the future, research on Bayesian formulations coupled with sequential testing is warranted to perform tests of significance for regularization and clustering methods which can indicate the optimal number of components and parameter settings.

Overall, our analysis provides an overall understanding about the gene regulation by CNA and their impact on several breast cancer related outcomes. In future, as larger studies emerge with a greater coverage and spectrum of molecular phenotypes, a more comprehensive insight as to the intermediate regulatory mechanism will take shape. For that, it will be imperative to move beyond identifying single variant-gene or variant-outcome associations and conceptualize associations in context of networks and modules.

Acknowledgement

DD was supported by R01 grant from the National Human Genome Research Institute [1 R01 HG010480-01; PI Dr. Nilanjan Chatterjee]. AS and JS were supported by R01 grant from the National Cancer Institute [7 R01 CA197402-05; PI: Jaya Satagopan].

References:

1. Beroukhim R, Mermel CH, Porter D, et al. The landscape of somatic copy-number alteration across human cancers. *Nature*. 2010;463(7283):899-905. doi:10.1038/nature08822
2. Nik-Zainal S, Davies H, Staaf J, et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*. 2016;534(7605):47-54. doi:10.1038/nature17676
3. Ciriello G, Gatza ML, Beck AH, et al. Comprehensive Molecular Portraits of Invasive Lobular Breast Cancer. *Cell*. 2015;163(2):506-519. doi:10.1016/j.cell.2015.09.033
4. Zhang H, Ahearn TU, Lecarpentier J, et al. Genome-wide association study identifies 32 novel breast cancer susceptibility loci from overall and subtype-specific analyses. *Nat Genet*. 2020. doi:10.1038/s41588-020-0609-2
5. Dawson S-J, Rueda OM, Aparicio S, Caldas C. A new genome-driven integrated classification of breast cancer and its implications. *EMBO J*. 2013;32(5):617-628. doi:10.1038/emboj.2013.19
6. Alvarez RH, Valero V, Hortobagyi GN. Emerging Targeted Therapies for Breast Cancer. *J Clin Oncol*. 2010;28(20):3366-3379. doi:10.1200/JCO.2009.25.4011
7. Stuart D, Sellers WR. Linking somatic genetic alterations in cancer to therapeutics. *Curr Opin Cell Biol*. 2009;21(2):304-310. doi:10.1016/j.ceb.2009.02.001
8. Curtis C, Shah SP, Chin S-F, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*. 2012;486(7403):346-352. doi:10.1038/nature10983
9. Zender L, Spector MS, Xue W, et al. Identification and Validation of Oncogenes in Liver Cancer Using an Integrative Oncogenomic Approach. *Cell*. 2006. doi:10.1016/j.cell.2006.05.030
10. Eder AM, Sui X, Rosen DG, et al. Atypical PKC contributes to poor prognosis through loss of apical-basal polarity and Cyclin E overexpression in ovarian cancer. *Proc Natl Acad Sci*. 2005;102(35):12519-12524. doi:10.1073/pnas.0505641102
11. Zhang L, Feizi N, Chi C, Hu P. Association analysis of somatic copy number alteration burden with breast cancer survival. *Front Genet*. 2018. doi:10.3389/fgene.2018.00421
12. Holland DG, Burleigh A, Git A, et al. ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium. *EMBO Mol Med*. 2011. doi:10.1002/emmm.201100122
13. Gatza ML, Silva GO, Parker JS, Fan C, Perou CM. An integrated genomics approach identifies drivers of proliferation in luminal-subtype human breast cancer. *Nat Genet*. 2014. doi:10.1038/ng.3073
14. Cai Y, Crowther J, Pastor T, et al. Loss of Chromosome 8p Governs Tumor Progression and Drug Response by Altering Lipid Metabolism. *Cancer Cell*. 2016. doi:10.1016/j.ccell.2016.04.003

15. Boyle EA, Li YI, Pritchard JK. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell*. 2017;169(7):1177-1186. doi:10.1016/j.cell.2017.05.038
16. Liu X, Li YI, Pritchard JK. Trans Effects on Gene Expression Can Drive Omnigenic Inheritance. *Cell*. 2019. doi:10.1016/j.cell.2019.04.014
17. Hardoon DR, Shawe-Taylor J. Sparse canonical correlation analysis. *Mach Learn*. 2011. doi:10.1007/s10994-010-5222-7
18. Witten DM, Tibshirani R, Hastie T. A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*. 2009;10(3):515-534. doi:10.1093/biostatistics/kxp008
19. Pereira B, Chin S-F, Rueda OM, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nat Commun*. 2016;7(1):11479. doi:10.1038/ncomms11479
20. MSK Data Catalog. Breast Invasive Carcinoma (TCGA, Firehose Legacy). Cbioportal.
21. Pillai NS, Meng XL. An unexpected encounter with cauchy and levy. *Ann Stat*. 2016. doi:10.1214/15-AOS1407
22. Liu Y, Chen S, Li Z, Morrison AC, Boerwinkle E, Lin X. ACAT: A Fast and Powerful p Value Combination Method for Rare-Variant Analysis in Sequencing Studies. *Am J Hum Genet*. 2019;104(3):410-421. doi:10.1016/j.ajhg.2019.01.002
23. Chen L, Zhou Y. A fast and powerful aggregated Cauchy association test for joint analysis of multiple phenotypes. *Genes Genomics*. 2021;43(1):69-77. doi:10.1007/s13258-020-01034-3
24. Moore JE, Purcaro MJ, Pratt HE, et al. Expanded encyclopaedias of DNA elements in the human and mouse genomes. *Nature*. 2020;583(7818):699-710. doi:10.1038/s41586-020-2493-4
25. Dunham I, Kundaje A, Aldred SF, et al. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012;489(7414):57-74. doi:10.1038/nature11247
26. Keenan AB, Torre D, Lachmann A, et al. ChEA3: transcription factor enrichment analysis by orthogonal omics integration. *Nucleic Acids Res*. 2019. doi:10.1093/nar/gkz446
27. Lachmann A, Xu H, Krishnan J, Berger SI, Mazloom AR, Ma'ayan A. ChEA: transcription factor regulation inferred from integrating genome-wide ChIP-X experiments. *Bioinformatics*. 2010;26(19):2438-2444. doi:10.1093/bioinformatics/btq466
28. Ikeda H, Taira N, Hara F, et al. The estrogen receptor influences microtubule-associated protein tau (MAPT) expression and the selective estrogen receptor inhibitor fulvestrant downregulates MAPT and increases the sensitivity to taxane in breast cancer cells. *Breast Cancer Res*. 2010;12(3):R43. doi:10.1186/bcr2598
29. Charan M, Das S, Mishra S, et al. Macrophage migration inhibitory factor inhibition as a novel

therapeutic approach against triple-negative breast cancer. *Cell Death Dis*. 2020;11(9):774. doi:10.1038/s41419-020-02992-y

30. Rodchenkov I, Babur O, Luna A, et al. Pathway Commons 2019 Update: integration, analysis and exploration of pathway data. *Nucleic Acids Res*. October 2019. doi:10.1093/nar/gkz946
31. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. *Cell Syst*. 2015;1(6):417-425. doi:10.1016/j.cels.2015.12.004
32. Dang C V. MYC on the Path to Cancer. *Cell*. 2012;149(1):22-35. doi:10.1016/j.cell.2012.03.003
33. Rosen EM, Fan S, Pestell RG, Goldberg ID. BRCA1 gene in breast cancer. *J Cell Physiol*. 2003. doi:10.1002/jcp.10257
34. Kuchenbaecker KB, Hopper JL, Barnes DR, et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA - J Am Med Assoc*. 2017. doi:10.1001/jama.2017.7112
35. Tai YC, Domchek S, Parmigiani G, Chen S. Breast cancer risk among male BRCA1 and BRCA2 mutation carriers. *J Natl Cancer Inst*. 2007. doi:10.1093/jnci/djm203
36. Dang H, Pomyen Y, Martin SP, et al. NELFE-Dependent MYC Signature Identifies a Unique Cancer Subtype in Hepatocellular Carcinoma. *Sci Rep*. 2019. doi:10.1038/s41598-019-39727-9
37. Han L, Zan Y, Huang C, Zhang S. NELFE promoted pancreatic cancer metastasis and the epithelial□to□mesenchymal transition by decreasing the stabilization of NDRG2 mRNA. *Int J Oncol*. October 2019. doi:10.3892/ijo.2019.4890
38. Sun C-C, Li S-J, Hu W, et al. Comprehensive Analysis of the Expression and Prognosis for E2Fs in Human Breast Cancer. *Mol Ther*. 2019;27(6):1153-1165. doi:10.1016/j.ymthe.2019.03.019
39. Khaleel SS, Andrews EH, Ung M, DiRenzo J, Cheng C. E2F4 regulatory program predicts patient survival prognosis in breast cancer. *Breast Cancer Res*. 2014. doi:10.1186/s13058-014-0486-7
40. Chhabra A, Fernando H, Watkins G, Mansel RE, Jiang WG. Expression of transcription factor CREB1 in human breast cancer and its correlation with prognosis. *Oncol Rep*. 2007. doi:10.3892/or.18.4.953
41. Fang Z, Lin A, Chen J, et al. CREB1 directly activates the transcription of ribonucleotide reductase small subunit M2 and promotes the aggressiveness of human colorectal cancer. *Oncotarget*. 2016. doi:10.18632/oncotarget.12938
42. Al-Dhaheri MH, Shah YM, Basrur V, Pind S, Rowan BG. Identification of novel proteins induced by estradiol, 4-hydroxytamoxifen and acolbifene in T47D breast cancer cells. *Steroids*. 2006. doi:10.1016/j.steroids.2006.07.006

43. Craig Jordan V. The role of tamoxifen in the treatment and prevention of breast cancer. *Curr Probl Cancer*. 1992;16(3):134-176. doi:10.1016/0147-0272(92)90002-6

44. Ahmad A, Zhang W, Wu M, Tan S, Zhu T. Tumor-suppressive miRNA-135a inhibits breast cancer cell proliferation by targeting ELK1 and ELK3 oncogenes. *Genes Genomics*. 2018;40(3):243-251. doi:10.1007/s13258-017-0624-6

45. Sarvagalla S, Kolapalli SP, Vallabhapurapu S. The Two Sides of YY1 in Cancer: A Friend and a Foe. *Front Oncol*. 2019;9. doi:10.3389/fonc.2019.01230

46. Chen L-F. Tumor suppressor function of RUNX3 in breast cancer. *J Cell Biochem*. 2012:n/a-n/a. doi:10.1002/jcb.24074

47. Ye DZ, Field J. PAK signaling in cancer. *Cell Logist*. 2012;2(2):105-116. doi:10.4161/cl.21882

48. Chang Y, Park KH, Lee JE, Han K-C. Phosphoproteomic analysis reveals PAK2 as a therapeutic target for lapatinib resistance in HER2-positive breast cancer cells. *Biochem Biophys Res Commun*. 2018;505(1):187-193. doi:10.1016/j.bbrc.2018.09.086

49. MIN W, LIU J, ZHANG S. Sparse Weighted Canonical Correlation Analysis. *Chinese J Electron*. 2018;27(3):459-466. doi:10.1049/cje.2017.08.004

50. Dutta D, He Y, Saha A, Arvanitis M, Battle A, Chatterjee N. Novel Aggregative trans-eQTL Association Analysis of Known Genetic Variants Detect Trait-specific Target Gene-sets. *medRxiv*. 2020:2020.09.29.20204388. https://doi.org/10.1101/2020.09.29.20204388.

Figure 1: Results from sCCA analysis (A) Chromosomal subregions identified by 14 CNA components. For each CNA component the region between the most distal CNA selected in that component is marked. (B) Average squared correlation between the CNA selected in CNA components and genes selected in Gene components. Correlation between Genes in components 2 and CNA in component 8 (and vice versa) are observed due possible correlation between selected CNA and long-range regulatory effects. Similar correlation is observed for components 5 & 7 as well.

Figure 2: **Examples of trans associations** identified in using the genes selected in (A) Gene component 1 (B) Gene component 4. Several genes in a chromosome different from that of the selected CNA is identified. Further, numerous distal genes (> 10Mb) on the same chromosome are detected as well.

Figure 3: Association analysis with breast cancer related outcomes. (A) DNALI1 gene on chromosome 1 associated with ER status and trans-regulated by a CNA of transcription factor CHD1 on chromosome 5. (B) Trans regulation of CD2BP2 gene on chromosome 16p11.2 by a CNA in ZNF263 gene located in chromosome 16q13.3 which are approximately 27Mb apart. (C) Association of CD2BP2 expression level with overall survival probability. Expression levels have been dichotomized as high and low using 75-th percentile as cut-off. (D) p-values of 72 genes identified to be strongly associated (p-value < 2.5×10^{-06}) with multiple outcomes, across the 7 outcomes. P-values < 1×10^{-12} are collapsed to 1×10^{-12} for the ease of viewing.

Figure 1

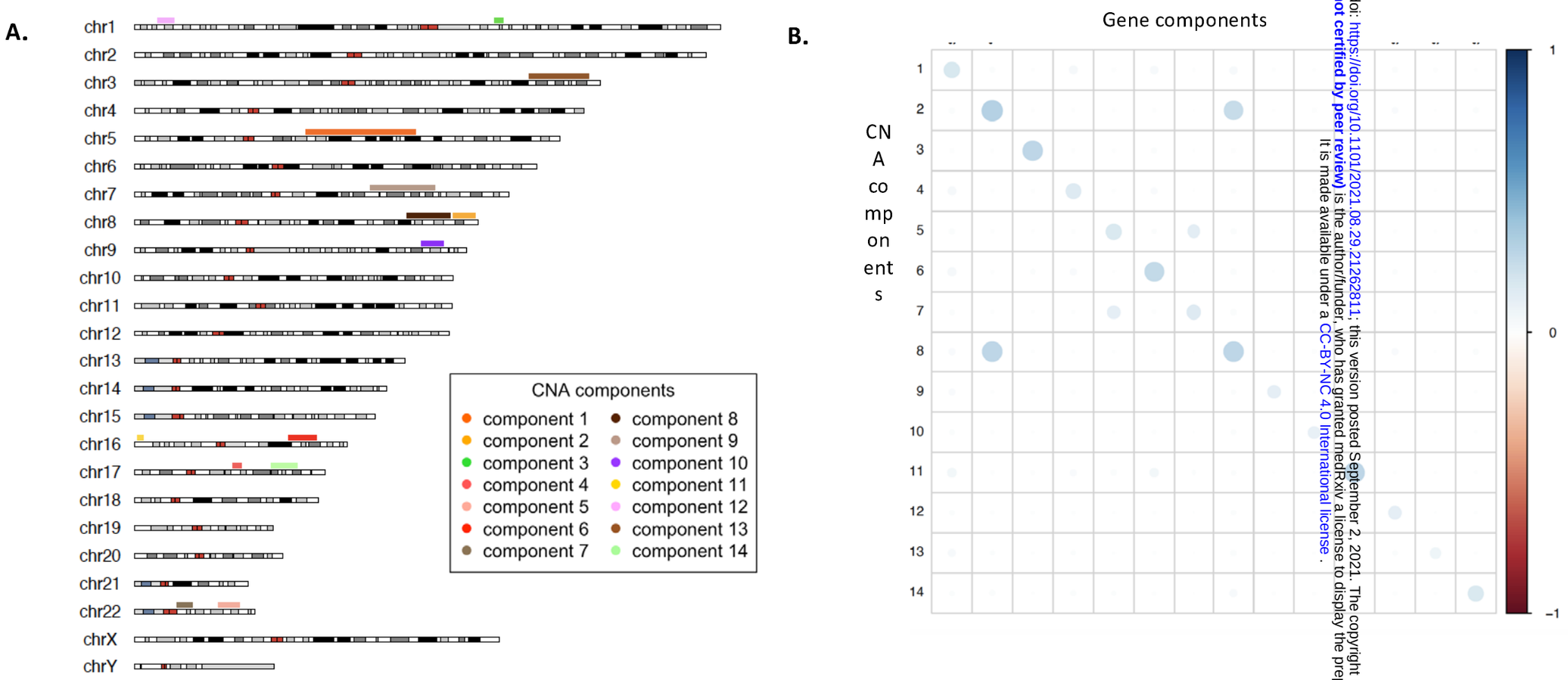
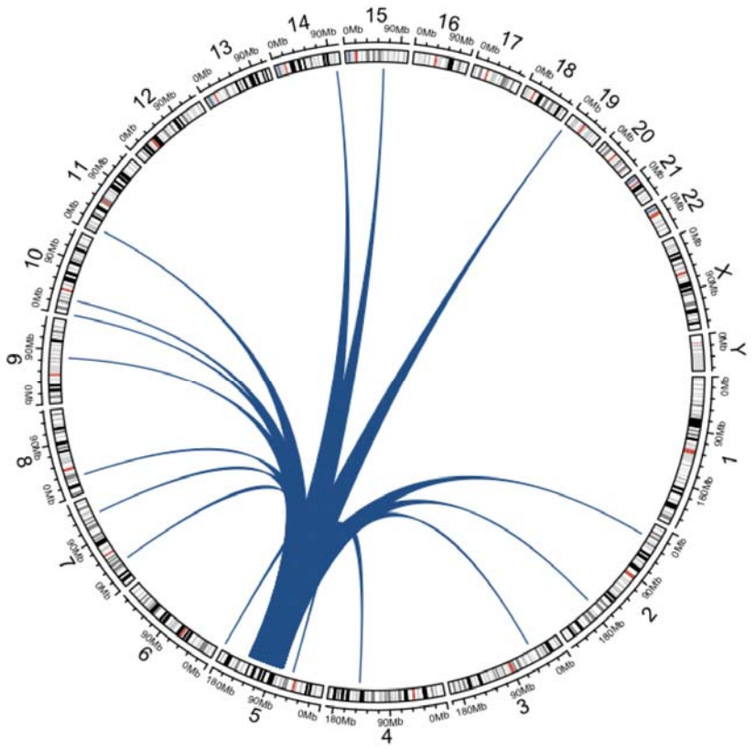


Figure 2

A.

Gene component 1



B.

Gene component 4

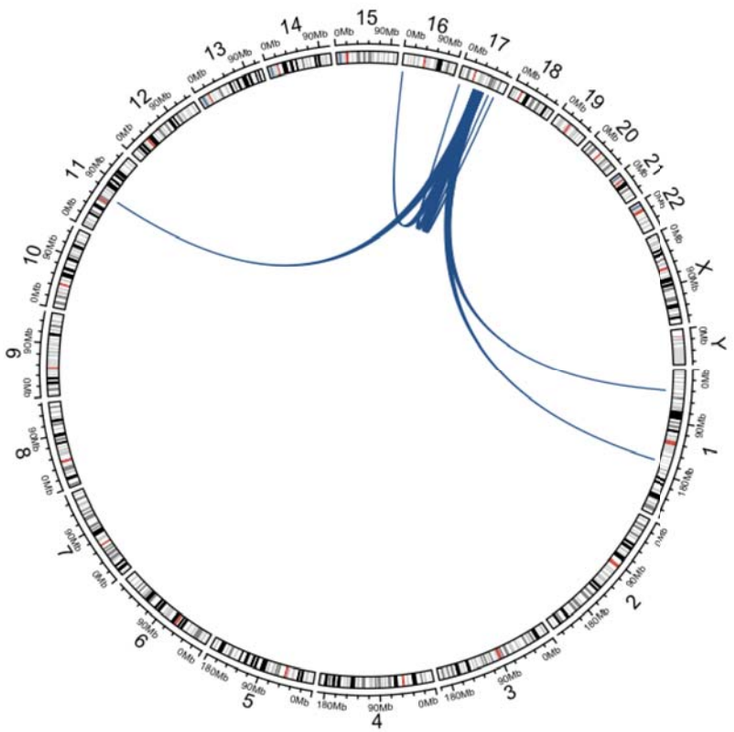


Figure 3

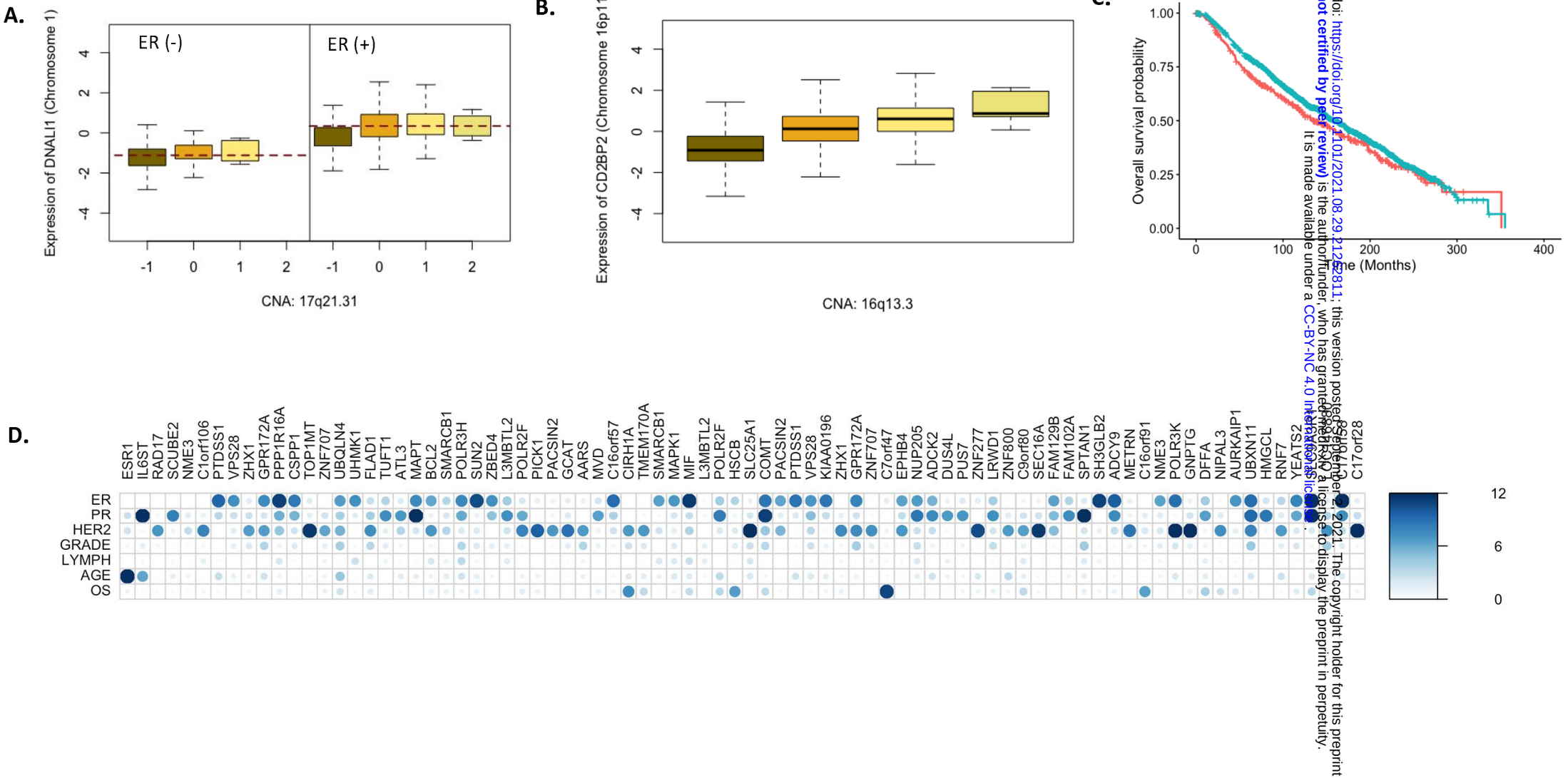


Table 1: Description of the 14 gene modules and CNA components identified through sCCA.

	CNA components			Gene components (modules)		
sCCA component	Number of CNA selected	Chromosome Number	Genomic location of selected CNA (Mb)	Number of Genes selected	Genes on different chromosome	Distal genes on same chromosome
1	139	5	72.17 – 119.54	70	24	18
2	107	8	135.45 – 145.01	59	0	14
3	131	1	153.20-156.96	56	0	6
4	115	17	41.86 – 45.50	67	9	4
5	145	22	35.64 – 44.57	63	0	6
6	130	16	65.28 – 77.25	58	0	6
7	125	22	17.95 – 24.63	58	0	10
8	125	8	115.95 – 134.48	59	0	24
9	145	7	100.40 – 127.61	71	0	17
10	151	9	121.82 – 131.49	74	0	1
11	129	16	1.06 – 3.72	64	1	11
12	117	1	9.65 – 16.64	63	0	12
13	147	3	168.00 – 193.44	71	0	11
14	145	17	58.16 – 69.14	62	0	1

Table 2: Description of the seven breast cancer related outcomes analyzed, and the number of genes associated significantly.

Outcome	% cases	Median survival	Significant Genes (FDR < 0.05)
Estrogen Receptor status (ER)	76.6	-	210
Progesterone receptor status (PR)	52.9	-	237
Human Epidermal growth factor Receptor 2 status (HER2)	12.4	-	255
Grade (Grade 3 vs Grade lower than 3)	47.5	-	65
Lymph Nodes Examined to be present (present vs absent)	47.8	-	12
Age at diagnosis (< 50 years)	78.4	-	100
Overall Survival	-	154.2	73

Table 3: Different categories of pathways enriched for the 210 genes associated (FDR < 0.05) with ER status.

Category	Pathway	Adjusted p-value	Genes in pathway	Genes overlap
GO	Cellular macromolecule localization	5.7×10^{-09}	1886	39
	Intracellular protein transport	1.2×10^{-07}	1156	28
	Cellular response to DNA damage stimulus	4.4×10^{-07}	841	23
	Catalytic complex	4.7×10^{-10}	1552	36
	Adenyl Nucleotide binding	3.9×10^{-06}	1536	30
Hallmark	Estrogen Response (early)	4.9×10^{-03}	200	7
	DNA repair	4.9×10^{-03}	149	6
	E2F targets	9.3×10^{-03}	200	6
	MYC targets	9.3×10^{-03}	200	6
	MTORC1 signaling	4.5×10^{-02}	200	5
Curated Gene sets	NIKOLSKY_BREAST_CANCER_8Q23_Q24_AMPLICON	8.2×10^{-14}	157	17
	PUJANA_BRCA1_PCC_NETWORK	2.3×10^{-08}	1617	34
	CHARAFE_BREAST_CANCER_LUMINAL_VS_MESENCHYMAL_UP	4.3×10^{-07}	446	17
	VANTVEER_BREAST_CANCER_ESR1_UP	5.7×10^{-05}	208	15
	PUJANA_BRCA2_PCC_NETWORK	5.4×10^{-04}	423	12
Immunologic Signatures	GSE4984_GALECTIN1_VS_VEHICLE_CTRL_TREATED_DC_DN	1.4×10^{-06}	198	12
	GSE2770_UNTREATED_VS_IL12_TREATED_ACT_CD4_TCELL_2H_DN	3.8×10^{-06}	200	12
	GSE19825_NAIVE_VS_IL2RAHIGH_DAY3_EFF_CD8_TCELL_DN	1.8×10^{-05}	200	11
TF targets	ENCODE: NELFE	6.5×10^{-46}	9442	173
	ENCODE: E2F4	2.8×10^{-34}	12626	180
	ENCODE: CREB1	5.7×10^{-33}	12289	177
	ChEA: EGR1	3.1×10^{-09}	5000	82
	ChEA: ELF3	2.2×10^{-07}	1760	40

Table 4: Different categories of pathways enriched for the 73 genes associated (FDR < 0.05) with overall survival (OS)

Category	Pathway	Adjusted p-value	Genes in pathway	Genes overlap
GO	RNA metabolic process	4.2×10^{-03}	1542	14
	Macromolecule catabolic process	4.2×10^{-03}	1366	13
	Cellular component disassembly	1.2×10^{-02}	537	8
Hallmark	Adipogenesis	2.8×10^{-02}	200	4
Curated Gene sets	DIAZ_CHRONIC_MEYLOGENOUS_LEUKEMIA_UP	1.2×10^{-03}	1397	14
	Reactome: Metabolism of RNA	1.3×10^{-03}	668	10
	NIKOLSKY_BREAST_CANCER_17Q21_Q25_AMPLICON	2.1×10^{-02}	332	6
Immunologic Signatures	GSE3982_NEUTROPHIL_VS_TH1_DN	4.8×10^{-04}	199	7
	GSE3982_EOSINOPHIL_VS_NKCELL_DN	1.5×10^{-03}	197	5
	GSE27786_NKCELL_VS_NEUTROPHIL_UP	8.1×10^{-03}	199	5
TF targets	ENCODE: RUNX3	8.1×10^{-16}	11816	63
	ENCODE: ELK1	1.5×10^{-14}	11349	61
	ENCODE: YY1	1.5×10^{-12}	12289	177
	ChEA: ETS1	8.0×10^{-08}	1359	20
	ChEA: PADI4	9.4×10^{-03}	877	9