Reframing Breast Cancer Molecular Subtypes: Modeling Expression Patterns with Orthogonal Dimensions

Michael J Madsen[1], Stacey Knight[1,2], Carol Sweeney[1], Rachel Factor[1], Mohamed Salama[1], Venkatesh Rajamanickam[1], Bryan E Welm[1], Sasi Arunachalam[1], Brandt Jones[1], Kerry Rowe[2], Melissa Cessna[2], Alun Thomas[1], Lawrence H. Kushi[3], Bette J Caan[3], Philip S Bernard[1], Nicola J Camp[1]

[1]School of Medicine, University of Utah, Salt Lake City, Utah, USA, 84112.
[2]Intermountain Healthcare, Salt Lake City, Utah, USA, 84107.
[3]Division of Research, Kaiser Permanente, Oakland, California, USA, 94612.

Abstract
Complex diseases can be highly heterogeneous. To characterize molecular heterogeneity, feature selection methods are often used to identify genes that capture key expression differences in the transcriptome. These gene sets are used in prediction algorithms to define distinct subtypes. Molecular subtyping has been used extensively in cancers and found to be informative for clinical care, e.g., PAM50 (Prediction Analysis of Microarray 50) for breast cancer. However, many tissues do not fit neatly into a single archetypal subtype. We propose that expression diversity can be more comprehensively represented with multiple quantitative dimensions, and that improved methods to model heterogeneity will generate new discoveries. Here, we apply principal components analysis to PAM50 gene expression data from 911 population-based breast tumors and identify orthogonal dimensions. These dimensions not only recapitulate categorical breast intrinsic subtypes, but also include dimensions not previously recognized. Furthermore, while 238 familial breast tumors (non-BRCA1/2 high-risk pedigrees) were not significantly enriched by intrinsic subtype, two novel expression dimensions were highly enriched in the pedigrees. Proof-of-concept gene-mapping using these dimensions identified a 0.5Mb genomewide significant region at 12q15 ($p=2.6\times10^{-8}$) segregating to 8 breast cancers through 32 meioses. The region contains *CNOT2*, a gene controlling cell viability via the CCR4-NOT transcriptional regulatory deadenylase complex. These findings suggest that the multiple dimension approach is a flexible and powerful method to characterize tissue expression within a defined feature set. Our results support the hypothesis that germline susceptibilities influence tumor characteristics and that expression dimensions partition genetic heterogeneity, providing new avenues for germline genetic studies.

Introduction
Many studies have shown that disease tissues exhibit molecular heterogeneity(1-3). In particular, characterization of tumor profiles has been used extensively to subtype cancers(4-6), with breast cancer being a quintessential example(7). While these discoveries have clearly advanced our knowledge of tissue molecular heterogeneity and led to opportunities for improved clinical care (8-12), a single categorical variable of mutually exclusive subtypes is inadequate to reflect the molecular complexity. We propose that, rather than deconstructing gene expression to one categorical subtypes, the definition of orthogonal dimensions is a more flexible approach that can maintain multiple important expression signals. This approach provides quantitative variables that can be used by subsequent studies to condition on molecular diversity. Many study designs

can benefit from improved molecular representation, ranging from clinical trials, to epidemiologic and genetic studies. Here, our particular application addresses the latter.

The high-risk pedigree design has been instrumental in the mapping and discovery of germline susceptibility genes, including for breast cancer(13, 14). Critical to success is an informative phenotype, and power is optimized for phenotypes where the underlying genetic heterogeneity is minimized. For example, in breast cancer, a focus on early onset disease led to evidence for autosomal dominant, high penetrance genes *BRCA1* and *BRCA2*(15, 16). However, beyond such early successes, little progress has been made with pedigree-based gene-mapping in the complex disease realm. The discovery of distinctive gene expression patterns(7) and breast tumor intrinsic subtypes (Luminal A, Luminal B, HER2-enriched, and Basal-like) in sporadic breast tumors illustrated different genetic paths to tumorigenesis and associations with clinical endpoints(8, 9). These findings also provided an opportunity to explore molecular subtypes as informative in familial breast cancer. A 'same-gene-same-molecular-subtype' hypothesis is supported by the fact that *BRCA1* tumors have distinct expression profiles(17) and are most often Basal-like(18). Further, small studies characterizing familial breast tumors (18 tumors in 8 families(19), and 23 tumors in 11 families(20)) observed tumor molecular subtype patterns consistent with an ability to partition non-*BRCA1/2* tumors. However, definitive evidence for this hypothesis and specific expression tumor signatures informative for family-based mapping remain to be defined. Here, we use the PAM50(21) to gene expression profile 238 breast tumors from 11 extended high-risk non-*BRCA1/2* Utah pedigrees. We compare these to 911 tumors representative of the population(22) to identify tumor expression dimensions in statistical excess in pedigrees. Under the 'same-gene-same-molecular-subtype' hypothesis these signatures will reflect an inherited susceptibility and be powerful for gene mapping.

Results

Novel breast tumor dimensions

The 50 gene features in the PAM50 assay were previously selected to represent major expression clusters in breast tumors. An associated classifier assigns each tumor to a single subtype based on proximity to archetypal subtype centroids (Luminal A, Luminal B, HER2-enriched, Basal-like, and Normal-like) (21). PAM50 expression data were available for 911 Caucasian breast tumors from the LACE and Pathways studies (LACE/Pathways)(22-24). We used principal component (PC) analysis to interpret the 50-dimensional gene expression space and identify tumor expression dimensions that explained the majority of variance. Using the PAM50 expression matrix from the LACE/Pathways tumor data (population-based), we defined five PCs explaining 68% of the total variance (30.5%, 18.9%, 10.2%, 5.3%, 3.2% explained by PCs 1-5 respectively). The relationship between PCs 1-5 and common intrinsic subtypes is illustrated in Figure 1 and gene coefficients (eigenvectors) are shown in Supplemental Table 1. PC1 represents ER signaling (major coefficients for *PGR*, *ESR1*, *NAT1*, *FOXA1*) and concurrently, in the opposing direction, proliferation (Kendall's tau: 0.65, Supplemental Figure 1). PC1 clearly delineates Basal-like from Luminal A tumors. PC2 includes strong coefficients for cytokeratins *KRT5*, *KRT14* and *KRT17*, as well as other basal markers (*SFRP1* and *MIA*) and differentiates Basal-like from Luminal B tumors. PC4 is the only component with a substantial coefficient for ERBB2, and contains substantial coefficients for growth factor genes (*EGFR*, *FGFR4* and *GRB7*). As expected, PC4 correlates well with the ERBB2 score (Kendall's tau: 0.55, Supplemental Figure 2), and differentiates HER2-enriched tumors. Together, PC1, PC2 and PC4

recapitulate all of the common intrinsic subtypes for which the PAM50 was developed (Figure 2).

Components PC3 and PC5 were novel: not associated with intrinsic subtypes (Figure 1b), nor highly correlated with PAM50 proliferation, PGR, ESR or ERBB2 scores. A notable characteristic for PC3 was that gene coefficients for basal cytokeratins were strong and that these were atypically co-expressed with ER-regulated genes. PC5 also exhibited coefficients indicating *KRT17* and ER-regulated gene expression in the same direction, but otherwise was most similar to PC4, including appreciable coefficients for *EGFR*, *FGFR4*, and *GRB7*.

We investigated these 5 PC PAM50 dimensions in RNA sequencing data for 745 breast tumors from Caucasian women in the TCGA project. Despite extensive differences in expression technology, all PCs were replicated in the TCGA data (Supplemental Figures 3 and 4) confirming the PC dimensions are robust and do not suffer unduly from over-fitting. In particular, the TCGA data replicated the novel dimensions PC3 and PC5. Using the whole transcriptome, we further interrogated genes beyond the PAM50 set to identify single genes with expression highly correlated with PC3 or PC5. No individual genes correlated highly with either PC3 or PC5, indicating that these novel dimensions are more likely representative of a combination of aberrantly expressed genes.

Enrichment of tumor dimensions PC3 and PC5 in high-risk pedigrees

Gene expression data was generated using the PAM50 RT-qPCR research assay(21) in 238 breast tumors from Caucasian women in 11 non-*BRCA1/2* high-risk pedigrees (example, Figure 3a,b). A previously described bioinformatics pipeline was used to assign intrinsic subtypes and provide gene scores for ESR1, PGR, ERBB2, and proliferation(25). Notably, pedigrees were not homogeneous by intrinsic subtype (Table 1, Figure 3b) and none were significantly increased for any subtype after controlling for multiple testing.

The five PCs established in the LACE/Pathways data were used to generate PC 1-5 scores in 238 breast tumors from women in high-risk breast cancer pedigrees. Significant differences in the quantitative PCs scores were identified between the pedigree and the population tumors (PC3 $p<1\times10^{-12}$; PC5 $p=7.6\times10^{-8}$). Furthermore, 6 pedigrees were individually significantly different from that expected in the general population for either PC3 and/or PC5 (Table 2). The significant elevation of PC3 and PC5 scores in high-risk pedigrees makes these excellent phenotypes to investigate for shared inherited susceptibility. For both dimensions, we defined tumors with scores above the 90th percentile as extreme.

Shared Genomic Segment (SGS) analysis

As proof-of-concept, we performed SGS gene mapping in Utah high-risk pedigree 1817, selected for significance of both PC3 and PC5. SGS is a genetic analysis technique to identify chromosomal regions inherited identical-by-descent across large numbers of meioses, specifically designed for high-density SNP array data and extended pedigrees(26).

Ten breast cancer cases from this pedigree had tumors extreme for PC3 (spanning multiple intrinsic subtypes: 5 Luminal A; 3 Luminal B and 1 HER2-enriched). Nine breast cancer cases had tumors extreme for PC5 (2 Luminal A, 4 Luminal B, 1 HER2-enriched and 2 Basal-like).

Four breast cancer cases were in both sets. Germline DNA for all 15 women was genotyped with the OmniExpress high-density single nucleotide polymorphism (SNP) assay and after quality control 571,489 SNPs were available for SGS analysis in 14 of the 15 selected women. The nine PC3-extreme breast cancer cases with SNP data were separated by 36 meioses, and the nine PC5-extreme breast cancer cases were separated by 43 meioses (Figure 3b).

SGS analysis was performed in pedigree 1817, separately for PC3-extreme and PC5-extreme tumors. One genomewide significant 0.5 Mb region (70.3-70.8 Mb, hg18) was identified for the PC3-extreme tumors at chromosome 12q15 ($p=2.6\times10^{-8}$, LOD equivalent=6.4). This locus was shared by 8 PC3-extreme breast cancer cases and inherited through 32 meioses. Only three genes reside in the SGS region: *CNOT2* (CCR4-NOT Transcription Complex Subunit 2), *KCNMB4* (Potassium Calcium-Activated Channel Subfamily M Regulatory Beta Subunit 4), and part of *MYRFL (*Myelin Regulatory Factor-Like). Notably, the gene *CNOT2* is a subunit of the CCR4-NOT complex, a global transcriptional regulator(27, 28) involved in cell growth and survival(29, 30). No genomewide significant regions were found for the PC5-extreme tumor dimension.

Post-hoc inspection of the PC3-extreme tumor breast cancer cases sharing the 12q15 region did not reveal any previously suggested characteristics that alternatively could have been used to identify this subset. Cases did not cluster in particular branches of the pedigree (Figure 3b), are not homogeneous for intrinsic subtype (as previously noted), and do not share similar ages at diagnosis.

Discussion

We used the PAM50 gene expression assay, initially designed for molecular subtyping(21), to identify PC expression dimensions. Two of 5 expression dimensions (PC3 and PC5) are previously unrecognized tumor characteristics, independent of intrinsic subtypes and other clinico-pathologic information. These new dimensions include key genes considered important in other intrinsic types, but with atypical expression directions. Luminal breast cancers are uniformly ER positive and express cytokeratins 8 and 18(31). Conversely, Basal-like tumors are ER negative and express cytokeratins 5, 14, 17(31, 32). Novel dimension PC3 contradicted gene expression profiles seen in prototypic subtypes in that ER-regulated genes and basal cytokeratin genes acted in the same direction; PC3-extreme (high) tumors were luminal tumors that also expressed basal cytokeratins. Basal cytokeratin expression can also come from myoepithelial cells comprising “normal” ducts, and this signature is evident in the Normal-like subtype (Figure 1a). Proliferative tumors do not usually exhibit normal stroma contamination and are therefore an uncharacteristic feature in Luminal B and HER2-enriched tumors, which were included in the PC3-extreme tumors. This suggests the basal cytokeratin expression comes from the ER positive tumor epithelial cells; however, further investigation is necessary to confirm this atypical luminal tumor expression of basal cytokeratins.

While comparisons of high-risk pedigree and population tumors did not support a ‘same-gene-same-intrinsic-subtype’ hypothesis, a ‘same-gene-same-tumor-dimension’ hypothesis was supported. Novel breast tumor dimensions defined by PC3 and PC5 were found to be significantly enriched in non-BRCA1/2 high-risk pedigrees providing evidence that these dimensions may be associated with germline risk susceptibility. This provides new potential to identify genetic susceptibilities, reducing heterogeneity and preserving statistical power.

Consistent with this hypothesis, we presented a proof-of-concept gene-mapping case-study using PC3-extreme tumors which identified a genomewide significant 0.5 Mb region at 12q15 that inherited to 8 PC3-extreme breast cancer cases across 32 meioses ($p=2.8\times10^{-8}$). Of the three genes residing in the 0.5 Mb region, *CNOT2* is a compelling candidate because of its role as a regulatory protein of the CCR4 (carbon catabolite repressor-4)-NOT (negative on TATA) complex, which functions as a master regulator of transcription, translation and mRNA stability(33, 34). CCR4-NOT and *CNOT2* have been demonstrated to function in the regulation of DNA damage response, cell cycle progression, DNA replication stress response, and control of cell viability(30, 35-37). A transcriptional module of *CNOT2* has also been correlated with heritable susceptibility of metastatic progression in a mouse model of breast cancer. Direct involvement of *CNOT2* in metastasis was demonstrated; knockdown of *CNOT2* enhanced and overexpression of *CNOT2* attenuated lung metastasis of mouse mammary tumor cells(29). An attractive possibility is that a germline risk modifies *CNOT2* expression, leading to dysregulation of mechanisms controlling cell growth and DNA damage repair. The natural next step will be to determine specific genetic variants in our 12q15 locus and assess effect on *CNOT2*.

Beyond our application in gene-mapping, a PC dimension approach to deconstructing gene expression may have utility in other domains. Principal components are orthogonal measures, providing independent variables for multi-variate modeling. This flexibility has the potential for increased power over a single variable categorical approach by allowing multiple expression dimensions to be modeled simultaneously. In particular, studies using the PAM50 expression for tumor characterization can immediately explore the PCs described here proving additional opportunities to identify novel clinical or therapeutic associations. Furthermore, the deeper appreciation of the gene expression dimensions of breast tumors may be useful in illuminating functionally important tumor pathways, or particular tumor evolutions.

In summary, we have decomposed PAM50 gene expression into 5 PC dimensions and identified two novel breast tumor dimensions with significant evidence for underlying genetic heritability. Based on one of these novel tumor dimensions, we mapped a genomewide significant breast cancer locus at 12q15 and present a compelling breast cancer susceptibility candidate gene, *CNOT2*. The strong statistical significance achieved by the mapping of this new breast cancer susceptibility locus harkens back to the era of pedigree gene-mapping successes. These novel tumor dimensions may, indeed, reduce germline genetic heterogeneity and hold promise for a new wave of susceptibility gene discovery in breast cancer. Furthermore, the appreciation of all five expression dimensions from the PAM50 assay lends additional informative variables that can be assessed immediately, with potential for new discoveries in clinical outcome studies. Moreover, in any disease domain where features have been selected from big-data “omics” experiments, application of multi-dimensional reduction methods can be used to identify quantitiative orthogonal dimensions and provide new opportunities to model tissue complexities.

Methods

<u>Identification, selection and ascertainment of high-risk breast pedigrees</u>

High-risk breast cancer pedigrees were identified in the Utah Population Database (UPDB) through record linkage of a 16-generation genealogy and statewide cancer records from the Utah Cancer Registry (UCR). High-risk pedigrees were defined based on a statistical excess of breast cancer cases ($p<0.05$). Pedigrees with fewer than 15 meioses between cases or known to be

attributable to *BRCA1/2* from previous Utah studies (screen positive or linked to chromosomes 17q21 or 13q13) were removed from consideration. Record linkage between the UPDB and electronic medical records in the University of Utah and Intermountain Healthcare systems allowed identification of tumor blocks and identified 25 high-risk pedigrees, each with a minimum of 15 tumors availability (Table 1). Matched tumor and grossly uninvolved (GU) formalin-fixed paraffin-embedded (FFPE) breast tissue blocks were retrieved for pathological review and acquisition of tumor punches (ideal: 4×1.5 mm punches; minimum 1 punch) and GU scrolls (ideal: 7×15 µm full face scrolls; minimum 4 scrolls) was performed. This resulted in 391 quality-controlled tissue samples obtained from the Intermountain BioRepository (N=354) and the University of Utah Department of Pathology (N=37). In parallel, living breast cancer cases within the 25 high-risk pedigrees were invited to participate, including a blood draw. Eleven high-risk pedigrees contained 245 of the tissue samples. These 11 pedigrees were the focus of this study. Nucleic acid extraction was performed as described previously(25). After quality control, 238 breast cancer cases had both quality controlled tumor RNA and germline DNA available (45 blood-derived, 1 from saliva, the remaining from GU breast tissue). All women were Caucasian. Ethical approvals for the study were governed by IRBs at the University of Utah and Intermountain Healthcare.

LACE and Pathways Studies

The Life After Cancer Epidemiology (LACE) and Pathways Studies are prospective cohort studies of breast cancer prognosis(23, 24). Briefly, in the LACE Study, women were enrolled at least 6 months after diagnosis with Stage I-IIIb breast cancer, with baseline data collection in 2000. In the Pathways Study, 4,505 women at all stages of breast cancer were enrolled between 2006 and 2013, on average about two months after diagnosis. In these studies, most or all women were diagnosed with breast cancer in the Kaiser Permanente Northern California healthcare system; a small proportion of women in the LACE Study were also enrolled in the state of Utah. Both cohorts were sampled as broadly representative of breast cancers in the general population and participants were enrolled without regard to family history of cancer. A stratified random sample from the combined LACE/Pathways cohorts was selected for acquisition of primary tumor punches from FFPE blocks(38). Common hormone receptor positive, HER2-negative breast cancers, by immunohistochemistry subtypes, were sampled at a lower frequency. In this study, only tumors from the 911 Caucasian women were selected to match the ethnicity in the Utah pedigrees. Survey weights were provided to address the decreased sampling rate of certain immunohistochemistry subtypes.

PAM50 gene expression and intrinsic subtype determination

The PAM50 breast intrinsic subtyping assay by RT-qPCR was performed by the Bernard Lab at the Huntsman Cancer Institute for both the LACE/Pathways and pedigree tumors, as previously described (citations 11-13, 26) For each tumor sample a calibrated log-expression ratio was produced for each gene, producing an expression matrix. Centroid-based algorithms were used to generate quantitative normalized subtype scores for each of the four clinical subtypes plus a subtype characteristic of normal tissue ("Normal-like"). These subtype scores represent the correlation with "prototypic" breast tumors for each of the subtypes. Each tumor was categorized according to its maximal subtype score. Each tumor was additionally assigned quantitative proliferation, progesterone receptor (PGR), estrogen receptor (ESR) and ERRB2 expression scores(25).

Principal Component (PC) Analysis

Expression in the LACE/Pathways breast tumors(38) was used to represent that expected in the general population. Principal components analysis was performed on the PAM50 expression matrix (N=911) incorporating survey weights to correct for the sampling strategy. We performed a weighted random sample with replacement to the correct size (N=911) and a PC analysis was performed using core packages of R version 3.1.1. This procedure was performed 10,000 times and the resulting PCs from each iteration were aligned as necessary, then averaged and centered. The first five PCs were found to account for 68% of the total variance, with components beyond PC5 explaining diminishing amounts of variance. We used the inflection point on the variance scree plot to select 5 PCs to include (Supplemental Figure 5).

Comparison of Principal Components to Intrinsic Subtypes and other characteristics

Stacked intrinsic-subtype-specific histograms were used to explore patterns between each quantitative PC and the categorical intrinsic subtypes (Figure 1). Kendall's tau coefficient was used to quantify the correlation of each PC to the PAM50 quantitative scores for proliferation, PGR, ESR and ERBB2 expression (Supplemental Figures 1 and 2). A 3-dimensional graph of PC1, PC2 and PC4 was used to illustrate approximate classification of the four intrinsic subtypes from these PCs (Figure 2).

Expression patterns in The Cancer Genome Atlas (TCGA) data

Standardized FPKM values for RNA sequencing transcriptome data and intrinsic subtype for 748 breast cancer cases in Caucasian women from the TCGA Breast Invasive Carcinoma project were downloaded from the National Cancer Institute GDC portal. The PC rotation matrix derived from the LACE/Pathways data was applied to log-transformed standardized FPKM expression values to establish scores for PCs 1-5. Stacked intrinsic-subtype-specific histograms for each PC were generated (Supplemental Figure 3).

To explore whether novel dimensions PC3 and PC5 were acting as proxies for other genes, correlations between the TCGA PC3 and PC5 scores and gene expression for each of the other genes in the transcriptome were evaluated using Pearson product moment (ρ). The distribution of these correlations was Gaussian with no outliers for both PC3 and PC5. The three most highly correlated genes with PC3 score were *KRT14* (ρ=0.57), *KRT17* (ρ=0.57) and *KRT5* (ρ=0.55), reflecting PAM50 genes which were among the highest-ranked coefficients in the PC3 eigenvector. The most highly correlated PAM50 gene with PC5 score was *MMP11* (ρ=0.76) which is the highest rank coefficient in the PC5 eigenvector.

Comparison of PCs: High-risk pedigree vs LACE/Pathways

To avoid the unrealistic assumption that all high-risk pedigrees share the same risk variants, each pedigree was also tested individually. Quantitative PC scores for all pedigree tumors together and for each individual pedigree were tested for difference from the LACE/Pathways cohort using a weighted t-test that accounted for the LACE/Pathways sampling weights. PC1 was bi-modal and a likelihood ratio test of proportions was implemented. A Bonferroni correction was applied to account for testing across 11 pedigrees (type 1 error, α=0.0045).

Proof-of-concept gene-mapping in Pedigree 1817: extreme-PC3 and -PC5 tumors

Extreme PC3 tumors and extreme PC5 tumors were significantly increased in pedigree 1817 (Table 2). Breast cancer cases were assigned to be "PC3-extreme" or "PC5-extreme" if their PC score was above the 90th percentile of the LACE/Pathways scores. This resulted in a total set of 15 women with tumors of interest: 4 women whose tumors were extreme for both PC3 and PC5; 6 women whose tumors were extreme for only PC3; and 5 women whose tumors were extreme for only PC5. Germline DNA was available for all 15 women, either from peripheral blood or from GU breast tissue, and these were genotyped using the OmniExpress high-density SNP array using standard Illumina protocols. Quality control included: duplicate check, sex check, SNP call-rate (95%), sample call rate (90%, more liberal due to the FFPE DNA), failure of Hardy-Weinberg equilibrium ($p \leq 1 \times 10^{-5}$). After quality control, 571,489 SNPs in 14 women were available for shared genomic segment (SGS) analysis.

SGS analysis identifies statistically significant chromosomal regions shared by multiple, distant relatives. It is based on evaluating identity-by-state (IBS) sharing at consecutive SNP loci, with segregation from a common ancestor implied if the observed sharing is significantly longer than expected by chance(26). The method was developed specifically for extended pedigrees, with power gained from the unlikely event that long segments are inherited across a large number of meioses by chance. Statistical significance for an SGS chromosomal region is determined empirically using a gene-drop approach. Genotype data from the 1000Genomes Project(39) are used to estimate a graphical model for linkage disequilibrium (LD)(40), providing a probability distribution of chromosome-wide haplotypes in the population. Pairs of haplotypes are randomly assigned to pedigree founders according to the haplotype distribution. Mendelian segregation and recombination are simulated to generate genotypes for all pedigree members. The Rutgers genetic map(41) is used for a genetic map for recombination, with interpolation based on physical base pair position for SNPs not represented. Once the gene-drop is complete, the simulated SNP genotypes for the individuals of interest are used to determine chance sharing. This is repeated tens of millions of times to estimate the significance of the observed sharing for all autosomes. We performed nested SGS analyses for a hierarchical set of cases, beginning with those with the two most extreme PC values and expanding to the full PC-extreme set. The distribution of $-\log_{10}$ p-values for all chromosomes (including all subsets) follows a gamma distribution. Based on the assumption that the vast majority of all sharing in a genome is under the null, we define the appropriate gamma distribution from our observed data. Genomewide significance thresholds indicating a false positive rate 0.05 (1 false positive per 20 genomes), that account for all multiple testing (subsets and all chromosomes), were derived using the theory of large deviations as described previously(42). For 1817, the genomewide significance thresholds for PC3 and PC5 were 5.0×10^{-7} and 5.9×10^{-7}, respectively.

Data availability

Pedigree expression data referenced in this study will be available in the Gene Expression Omnibus database. All other relevant data are available from the authors upon request.

Computer code

The software used for SGS analysis is available at https://gitlab.com/camplab/sgs and https://gitlab.com/camplab/jps.

Acknowledgements
This study was supported by funding from National Cancer Institute grants R01 CA163353 (NJ Camp), R01 CA129059 (BJ Caan), R01 CA105274 (LH Kushi), and U01 CA195565 (LH Kushi/C Ambrosone). The work was also supported in part by the Genomics Core Facility and through the computational resources and staff expertise provided by the Center for High Performance Computing at the University of Utah. We thank the Pedigree and Population Resource of the Huntsman Cancer Institute, University of Utah (funded in part by the Huntsman Cancer Foundation) for its role in the ongoing collection, maintenance and support of the UPDB. We also acknowledge partial support for the UPDB through grant P30 CA2014 from the National Cancer Institute, University of Utah and from the University of Utah's Program in Personalized Health and Center for Clinical and Translational Science. The Utah Cancer Registry is funded by the National Cancer Institute's SEER Program, Contract No. HHSN261201300017I, with additional support from the Utah Department of Health and the University of Utah. Finally, we thank the participants and their families who make this research possible.

Author Contributions
The study and analysis procedures were designed by M.J.M., S.K., A.T, P.S.B. and N.J.C. LACE/Pathways cohort data was generated and shared by C.S., L.H.K and B.J.C. Tumor identification, acquisition and review was performed by R.F., M.S., K.R. and M.C. Pedigree sample processing, quality control and data analysis was performed by M.J.M., B.J. and V.R. Data interpretation was provided M.J.M., B.E.W., S.A., P.S.B. and N.J.C. The manuscript was primarily written by M.J.M. and N.J.C. with revisions and contributions by all authors.

Competing interests
Phil Bernard is an inventor of the PAM50 gene expression signature and a stakeholder in BioClassifier LLC, a company that licensed the PAM50 know-how to Nanostring Inc for the commercialization of Prosigna.

References

1. Kim B, *et al.* (2016) Gene expression profiles of human subcutaneous and visceral adipose-derived stem cells. *Cell Biochem Funct* 34(8):563-571.
2. Gimenez-Arnau A, *et al.* (2017) Transcriptome analysis of severely active chronic spontaneous urticaria shows an overall immunological skin involvement. *Allergy*.
3. Planken A, *et al.* (2017) Looking beyond the brain to improve the pathogenic understanding of Parkinson's disease: implications of whole transcriptome profiling of Patients' skin. *BMC Neurol* 17(1):6.
4. Cancer Genome Atlas Research N (2011) Integrated genomic analyses of ovarian carcinoma. *Nature* 474(7353):609-615.
5. Konstantinopoulos PA, Spentzos D, & Cannistra SA (2008) Gene-expression profiling in epithelial ovarian cancer. *Nat Clin Pract Oncol* 5(10):577-587.
6. Lapointe J, *et al.* (2004) Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc Natl Acad Sci U S A* 101(3):811-816.
7. Perou CM, *et al.* (2000) Molecular portraits of human breast tumours. *Nature* 406(6797):747-752.
8. Sorlie T, *et al.* (2001) Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A* 98(19):10869-10874.
9. van 't Veer LJ, *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415(6871):530-536.
10. Prat A, *et al.* (2016) Prognostic Value of Intrinsic Subtypes in Hormone Receptor-Positive Metastatic Breast Cancer Treated With Letrozole With or Without Lapatinib. *JAMA Oncol* 2(10):1287-1294.
11. Kommoss S, *et al.* (2017) Bevacizumab May Differentially Improve Ovarian Cancer Outcome in Patients with Proliferative and Mesenchymal Molecular Subtypes. *Clin Cancer Res*.
12. Seiler R, *et al.* (2017) Impact of Molecular Subtypes in Muscle-invasive Bladder Cancer on Predicting Response and Survival after Neoadjuvant Chemotherapy. *Eur Urol.*
13. Hall JM, *et al.* (1990) Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* 250(4988):1684-1689.
14. Wooster R, *et al.* (1994) Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* 265(5181):2088-2090.
15. Miki Y, *et al.* (1994) A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* 266(5182):66-71.
16. Tavtigian SV, *et al.* (1996) The complete BRCA2 gene and mutations in chromosome 13q-linked kindreds. *Nat Genet* 12(3):333-337.
17. Hedenfalk I, *et al.* (2001) Gene-expression profiles in hereditary breast cancer. *N Engl J Med* 344(8):539-548.
18. Sorlie T, *et al.* (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A* 100(14):8418-8423.
19. Hedenfalk I, *et al.* (2003) Molecular classification of familial non-BRCA1/BRCA2 breast cancer. *Proc Natl Acad Sci U S A* 100(5):2532-2537.
20. Larsen MJ, *et al.* (2014) RNA profiling reveals familial aggregation of molecular subtypes in non-BRCA1/2 breast cancer families. *BMC Med Genomics* 7:9.

21. Parker JS, *et al.* (2009) Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol* 27(8):1160-1167.
22. Caan BJ, *et al.* (2014) Intrinsic subtypes from the PAM50 gene expression assay in a population-based breast cancer survivor cohort: prognostication of short- and long-term outcomes. *Cancer Epidemiol Biomarkers Prev* 23(5):725-734.
23. Caan B, *et al.* (2005) Life After Cancer Epidemiology (LACE) Study: a cohort of early stage breast cancer survivors (United States). *Cancer Causes Control* 16(5):545-556.
24. Kwan ML, *et al.* (2008) The Pathways Study: a prospective study of breast cancer survivorship within Kaiser Permanente Northern California. *Cancer Causes Control* 19(10):1065-1076.
25. Bastien RR, *et al.* (2012) PAM50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. *BMC Med Genomics* 5:44.
26. Thomas A, Camp NJ, Farnham JM, Allen-Brady K, & Cannon-Albright LA (2008) Shared genomic segment analysis. Mapping disease predisposition genes in extended pedigrees using SNP genotype assays. *Ann Hum Genet* 72(Pt 2):279-287.
27. Collart MA (2003) Global control of gene expression in yeast by the Ccr4-Not complex. *Gene* 313:1-16.
28. Laribee RN, *et al.* (2007) CCR4/NOT complex associates with the proteasome and regulates histone methylation. *Proc Natl Acad Sci U S A* 104(14):5836-5841.
29. Faraji F, *et al.* (2014) An integrated systems genetics screen reveals the transcriptional structure of inherited predisposition to metastatic disease. *Genome Res* 24(2):227-240.
30. Ito K, *et al.* (2011) CNOT2 depletion disrupts and inhibits the CCR4-NOT deadenylase complex and induces apoptotic cell death. *Genes Cells* 16(4):368-379.
31. Abd El-Rehim DM, *et al.* (2004) Expression of luminal and basal cytokeratins in human breast carcinoma. *J Pathol* 203(2):661-671.
32. Nielsen TO, *et al.* (2004) Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res* 10(16):5367-5374.
33. Collart MA (2016) The Ccr4-Not complex is a key regulator of eukaryotic gene expression. *Wiley Interdiscip Rev RNA* 7(4):438-454.
34. Shirai YT, Suzuki T, Morita M, Takahashi A, & Yamamoto T (2014) Multifunctional roles of the mammalian CCR4-NOT complex in physiological phenomena. *Front Genet* 5:286.
35. Mulder KW, Winkler GS, & Timmers HT (2005) DNA damage and replication stress induced transcription of RNR genes is dependent on the Ccr4-Not complex. *Nucleic Acids Res* 33(19):6384-6392.
36. Rodriguez-Gil A, *et al.* (2016) HIPK family kinases bind and regulate the function of the CCR4-NOT complex. *Mol Biol Cell* 27(12):1969-1980.
37. Westmoreland TJ, *et al.* (2004) Cell cycle progression in G1 and S phases is CCR4 dependent following ionizing radiation or replication stress in Saccharomyces cerevisiae. *Eukaryot Cell* 3(2):430-446.
38. Sweeney C, *et al.* (2014) Intrinsic subtypes from PAM50 gene expression assay in a population-based breast cancer cohort: differences by age, race, and tumor characteristics. *Cancer Epidemiol Biomarkers Prev* 23(5):714-724.
39. Genomes Project C, *et al.* (2015) A global reference for human genetic variation. *Nature* 526(7571):68-74.

40. Abel HJ & Thomas A (2011) Accuracy and computational efficiency of a graphical modeling approach to linkage disequilibrium estimation. *Stat Appl Genet Mol Biol* 10:Article 5.
41. Matise TC, *et al.* (2007) A second-generation combined linkage physical map of the human genome. *Genome Res* 17(12):1783-1786.
42. Lander E & Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 11(3):241-247.

Tables

Table 1: Summary of the 11 high-risk Utah pedigrees

				PAM50 Intrinsic Subtype				
Pedigree	Familial risk p-value	Female BrCa	Tumor*	Basal-like	HER2-enriched	Luminal B	Luminal A	Normal-like
1800	0.00030	66	20	1	5	6	6	2
1801	0.04390	57	17	0	4	4	9	0
1808	0.03432	112	24	4	2	6	12	0
1809	0.04140	50	15	2	6	2	4	1
1812	0.01615	43	17	2	1	6	7	1
1817	0.00709	138	35	4	4	10	15	2
1818	0.00786	111	20	2	2	2	12	2
1819	0.04812	114	26	2	2	7	15	0
1820	0.01195	68	20	2	3	6	9	0
1821	0.01808	81	18	4	1	3	9	1
1822	0.00909	159	31	1	5	9	15	2

**Tumor samples with PAM50 data that passed QC. Three tumors belonged to two pedigrees and one tumor belonged to three pedigrees.*

Table 2: Results of weighted t-test of means for PC3 and PC5

		PC3		PC5	
Pedigree	n	p	t	p	t
1800	20	ns	1.74	ns	1.609
1801	17	ns	2.03	ns	2.311
1808	24	0.0008	4.76	0.0129	3.674
1809	15	0.0778	3.13	ns	2.468
1812	17	ns	2.12	ns	0.305
1817	35	0.0006	4.55	0.0133	3.508
1818	20	0.0147	3.72	ns	0.230
1819	26	0.0001	5.75	ns	1.437
1820	20	ns	1.60	0.0033	4.352
1821	18	0.1014	2.92	ns	1.735
1822	31	0.00004	5.54	ns	0.184

Comparing to the population (LACE/Pathways). Bonferroni corrected.

Figures

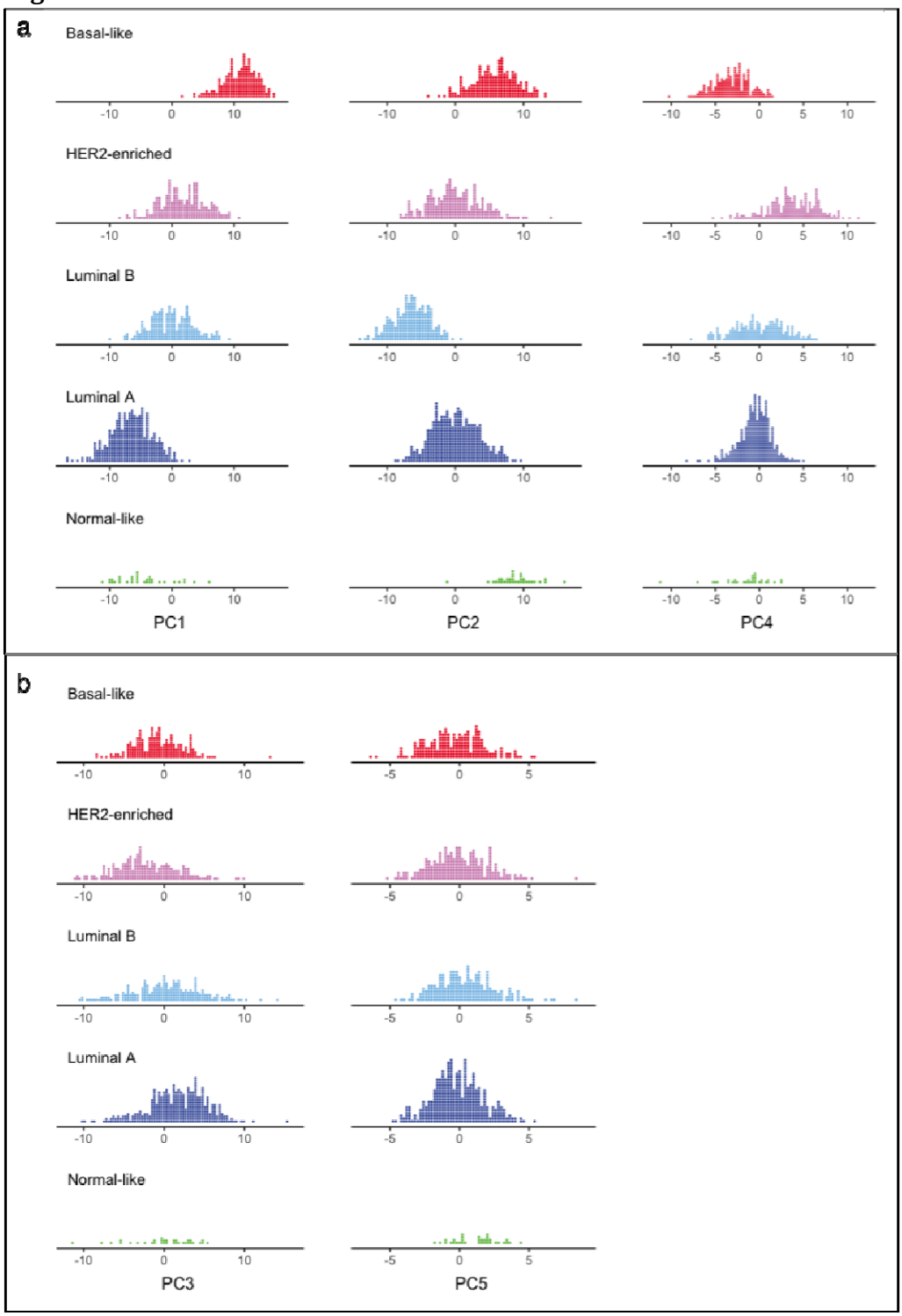


Figure 1. *Distribution of PC scores by intrinsic subtype in LACE/Pathways.* ***a.*** *PC1, PC2 and PC4 capture key features of intrinsic subtypes. PC1 illustrates proliferation and ER signaling, best differentiating Basal-like from Luminal A tumors. PC2 includes a strong signal from basal cytokeratins that clearly differentiates Basal-like from Luminal B. PC4 includes a strong signal from ERBB2 and differentiates HER2-enriched tumors. Together these 3 dimensions can recapitulate intrinsic subtype clusters.* ***b.*** *PC3 and PC5 are novel and are not associated with intrinsic subtype.*

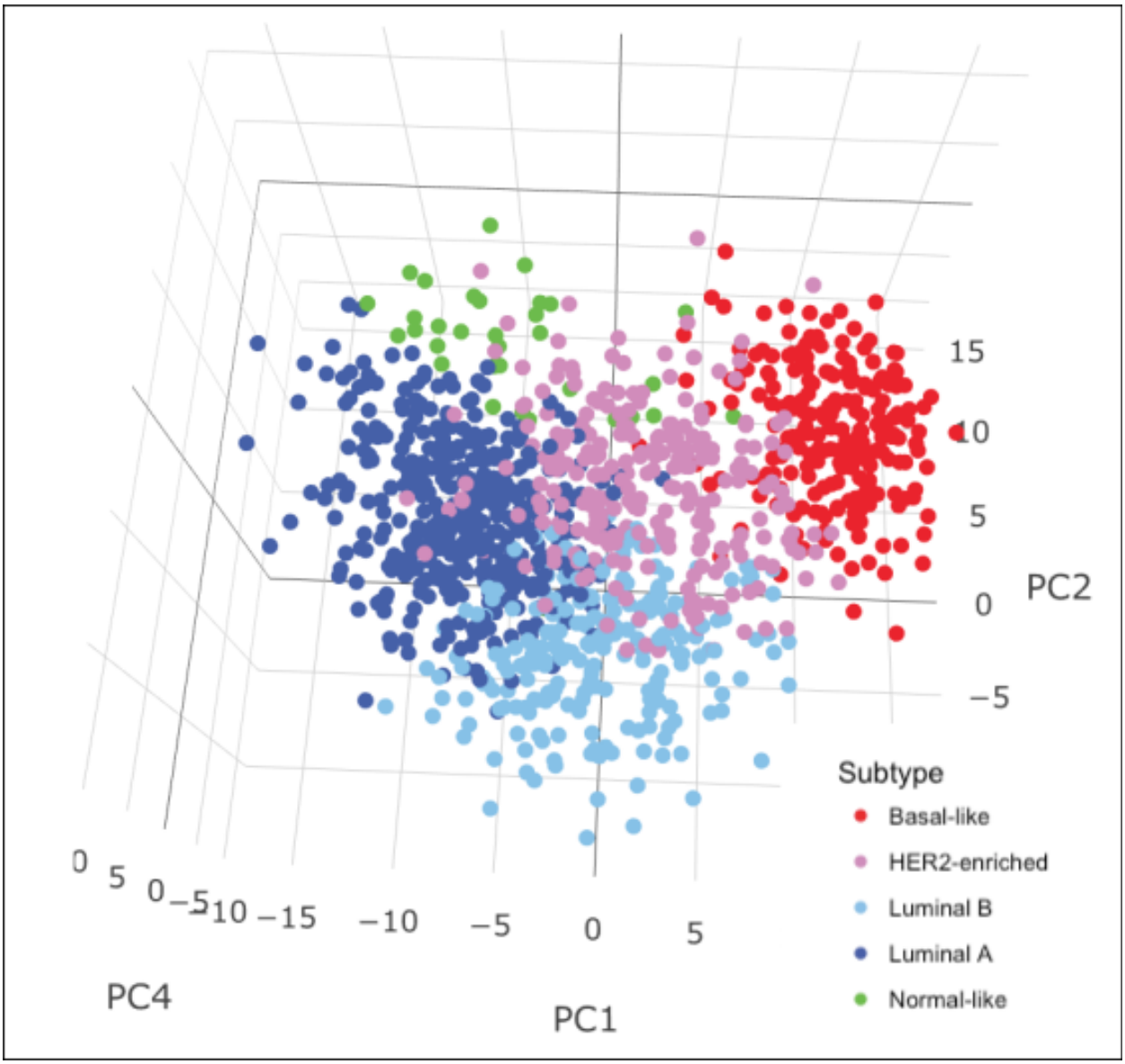


Figure 2. *PC1, PC2, and PC4 combined recapitulate PAM50 intrinsic subtypes*

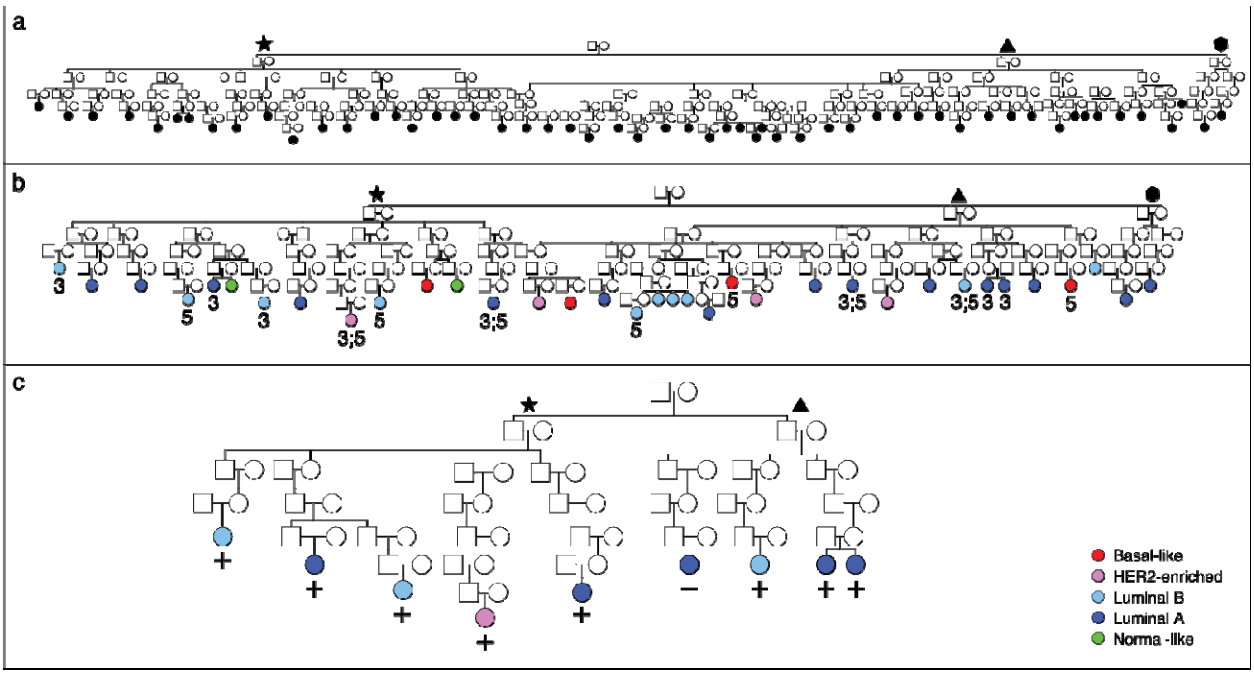
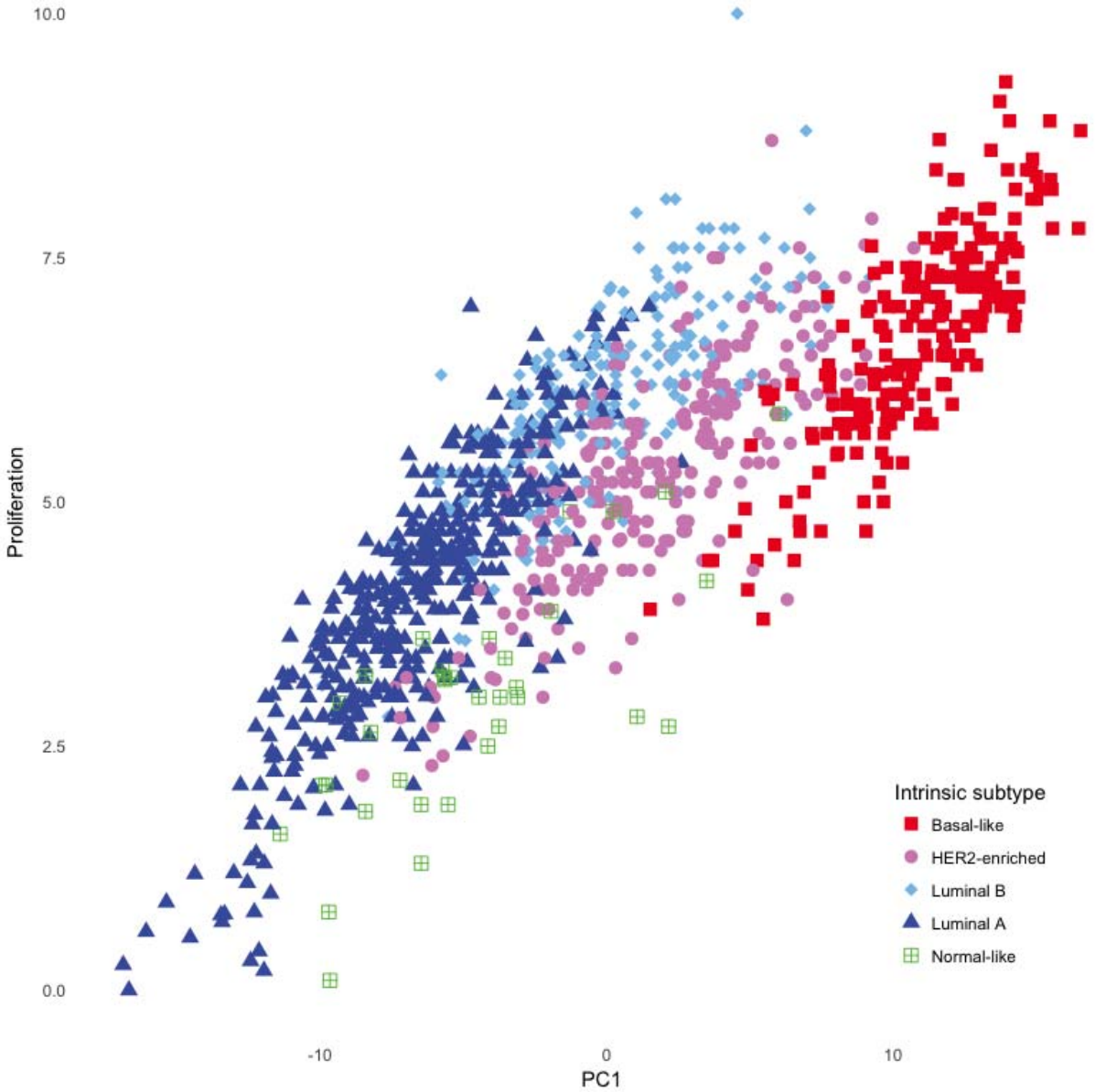
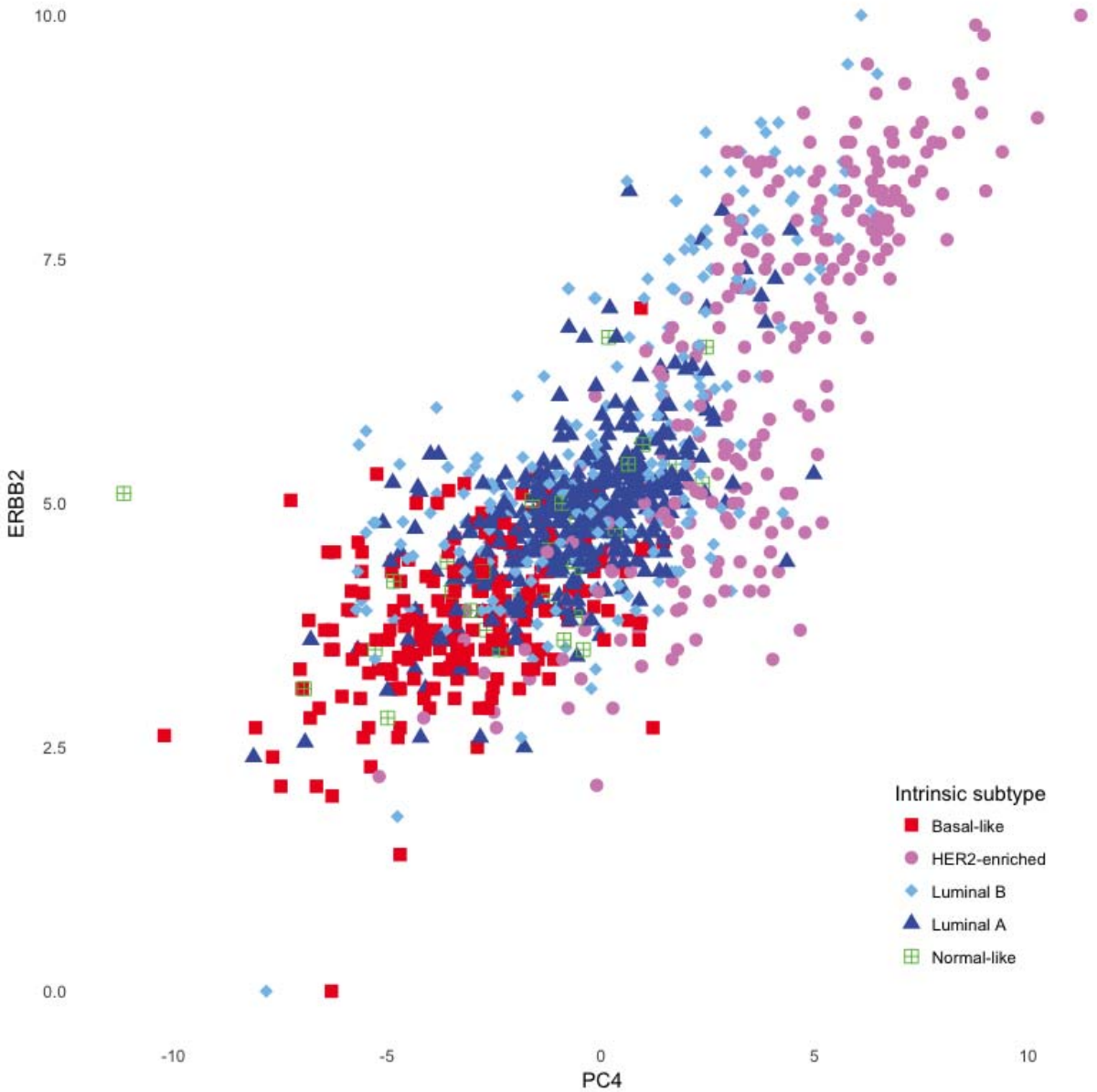


Figure* 3.** *Example Utah high-risk breast cancer pedigree 1817.* ***a. *Confirmed breast cancer cases are indicated in black. Star, triangle and hexagon symbols indicate pedigree branches.* ***b****. shows only those cases from (a) with tumor expression data available and indicates PAM50 intrinsic subtype by color. Cases whose tumors are extreme for PC3 are indicated by '3'; extreme for PC5 are indicated by '5'.* ***c.*** *shows only the PC3-extreme cases from (b). A '+' indicates those cases that share the genomewide significant region at 12q15.*

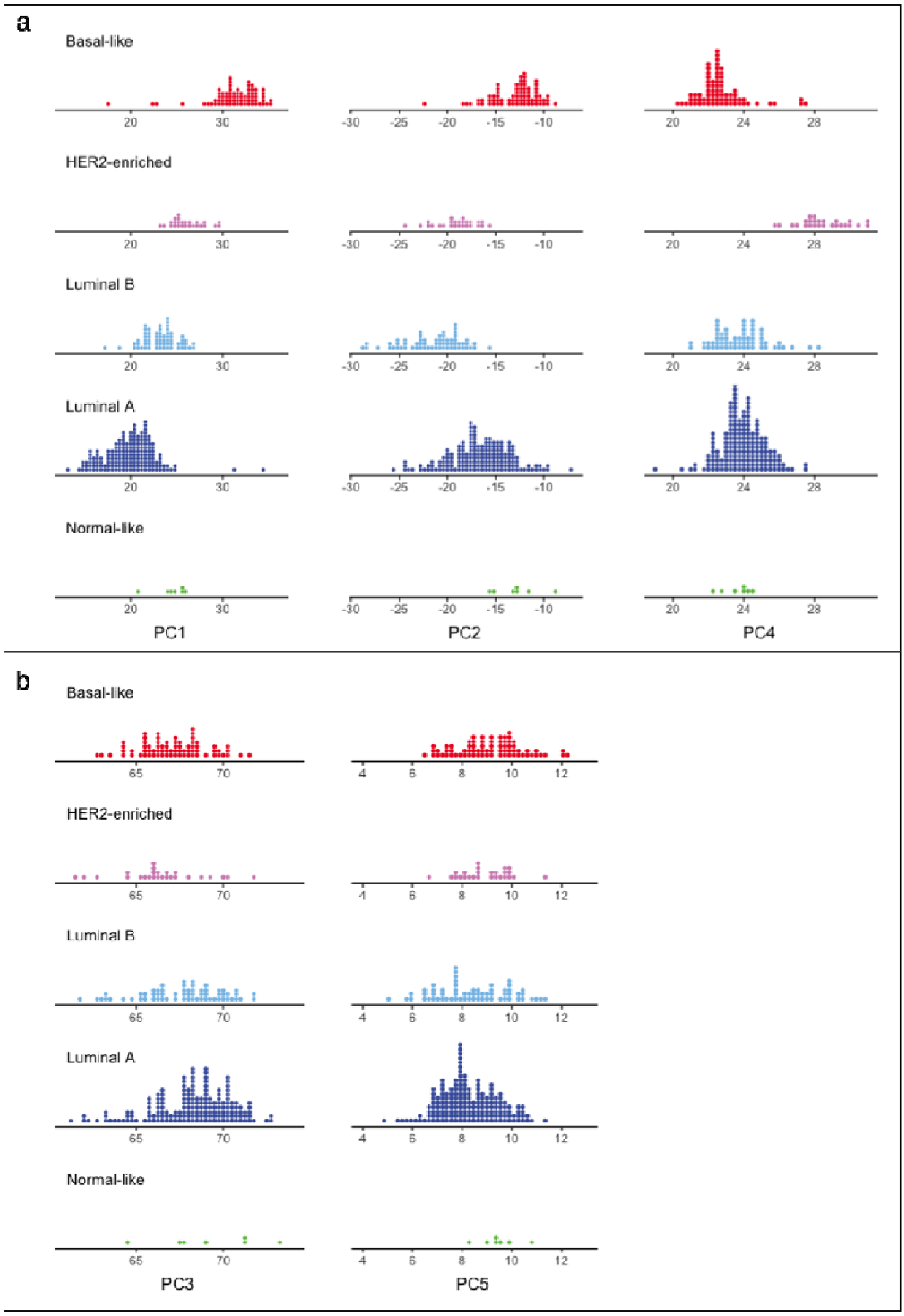
Supplemental Figures



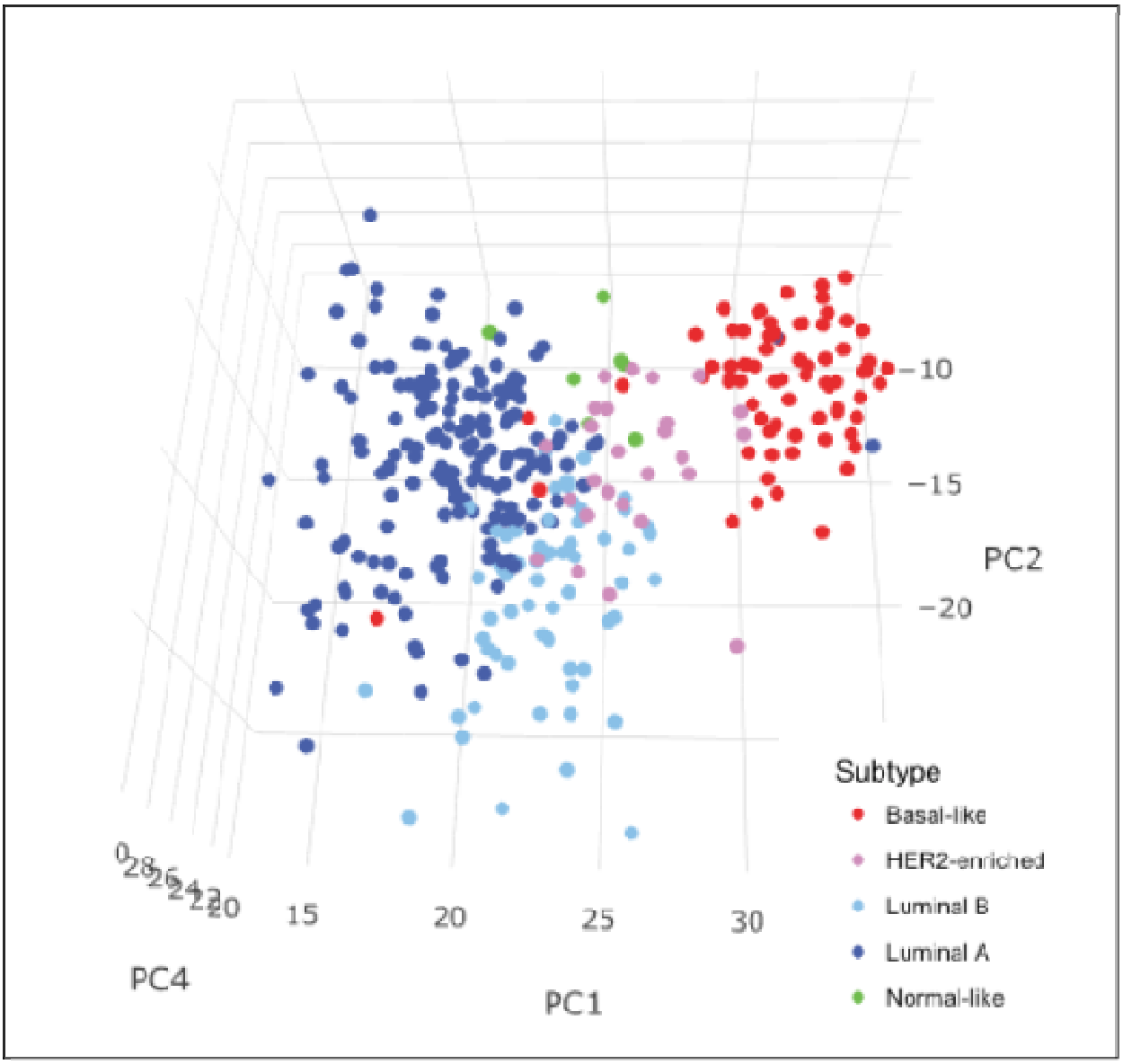
Supplemental Figure 1. *Correlation of PAM50 proliferation score with PC1 in LACE/Pathways.*



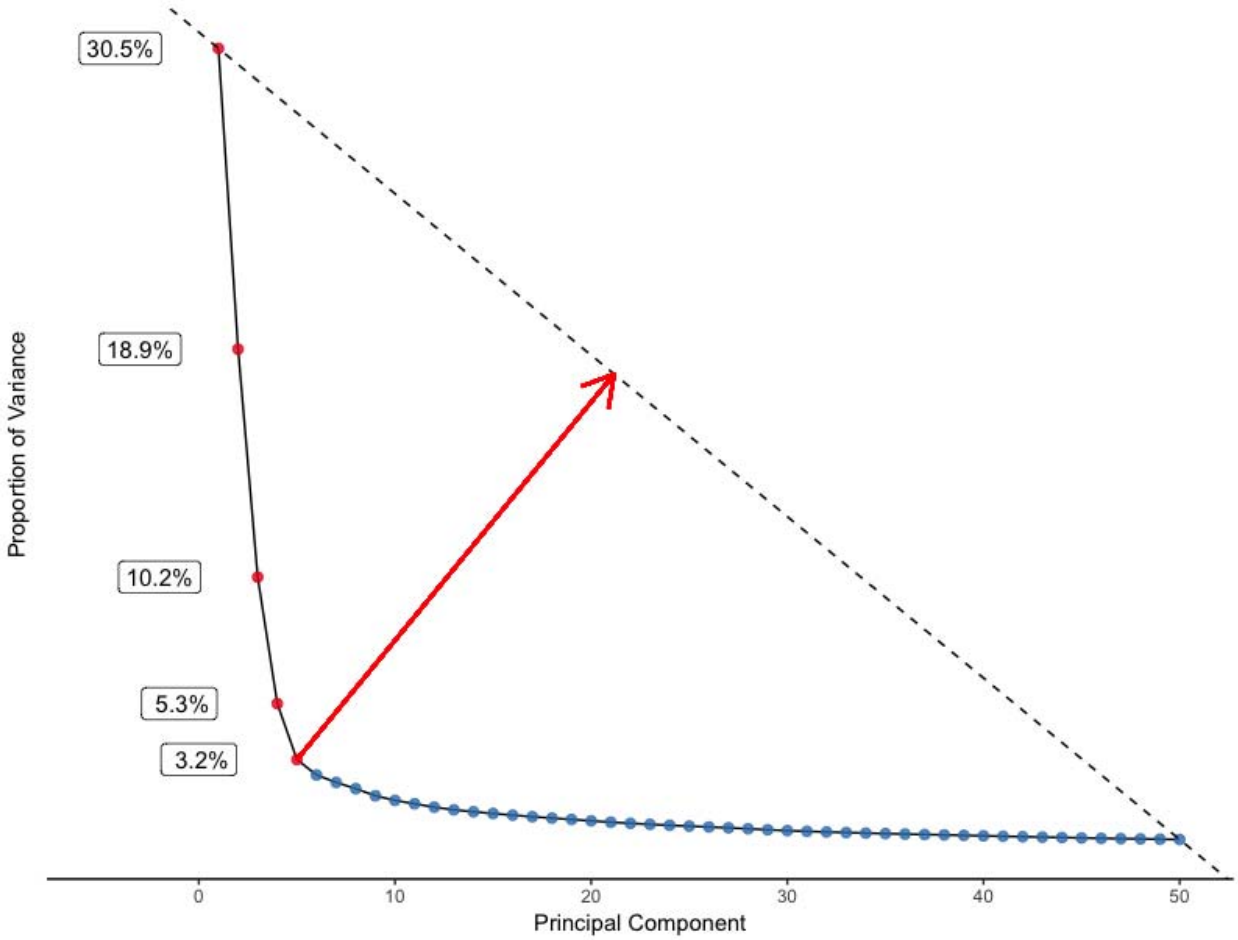
Supplemental Figure 2. *Correlation of PAM50 ERBB2 score with PC4 in LACE/Pathways.*



Supplemental Figure 3. *Distribution of PC scores by intrinsic subtype replicate in TCGA expression data.* ***a.*** *PC1, PC2 and PC4 capture key features of intrinsic subtypes.* ***b***. *PC3 and PC5 are independent of intrinsic subtype.*



Supplemental Figure 4*. Intrinsic subtypes remain evident in TCGA breast tumors based on direct application of the PC1, PC2, and PC4 equations (linear combinations of expression) derived from PAM50 (LACE/Pathways) to RNAseq data from TCGA breast tumors.*



Supplemental Figure 5. *Illustration of our Principal Component Analysis. Selection of PCs for study were made according to the inflection point in a scree plot of the proportion of sample variance explained by each PC.*