2017-06-01

Integrative Analysis to Evaluate Similarity Between BRCAness Tumors and BRCA Tumors

Weston Reed Bodily
Brigham Young University

Integrative Analysis to Evaluate Similarity Between

BRCAness Tumors and BRCA Tumors

Weston Reed Bodily

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Science

Stephen Piccolo, Chair
Marc Hansen
Perry Ridge

Department of Biology

Brigham Young University

ABSTRACT

Integrative Analysis to Evaluate Similarity Between BRCAness Tumors and BRCA Tumors

Weston Reed Bodily
Department of Biology, BYU
Master of Science

The term "BRCAness" is used to describe breast-cancer patients who lack a germline mutation in BRCA1 or BRCA2, yet who are believed to express characteristics similar to patients who do have a germline mutation in BRCA1 or BRCA2. Although it is hypothesized that BRCAness is related to deficiency in the homologous recombination repair (HRR) pathways, relatively little is understood about what drives BRCAness or what criteria should be used to assign patients to this category. We hypothesized that patients whose tumor carries a genomic or epigenomic aberration in BRCA1 or BRCA2 should be classified under the BRCAness category and that these tumors would exhibit downstream effects (additional mutations or gene-expression changes) similar to patients with germline BRCA1/2 mutations. To better understand BRCAness, we examined similarities and differences in gene-expression profiles and somatic-mutation "signatures" among 1054 breast-cancer patients from The Cancer Genome Atlas. First, we categorized patients into three categories: those who carried a germline BRCA1/2 mutation, those whose tumor carried a genomic aberration or DNA hypermethylation in BRCA1/2 (the BRCAness group), and those who fell into neither of the first two groups. Upon evaluating the gene-expression data in context of the PAM50 subtypes, we did not observe significant similarity between the germline BRCA1/2 and BRCAness groups, but we did observe enrichment within the basal subtype, especially for BRCAness tumors with hypermethylation of BRCA1/2. However, the gene-expression profiles were fairly heterogeneous; for example, BRCA1 patients differed significantly from BRCA2 patients. In agreement with prior findings, certain mutational signatures—especially "Signature 3"—were enriched for patients with germline BRCA1/2 mutations as well as for BRCAness patients. Furthermore, we observed significant similarity between germline BRCA1/2 patients and patients with germline mutations in PALB2, RAD51B, and RAD51C, genes that are key parts of the HRR pathway and that interact with BRCA1/2. Our findings suggest that the BRCAness category does have biological and clinical relevance but that the criteria for including patients in this category should be carefully defined, potentially including BRCA1/2 hypermethylation and homozygous deletions as well as germline mutations in PALB2, RAD51B, and RAD51C.

Keywords: breast cancer, multiomic analysis, BRCAness, BRCA1, BRCA2

ACKNOWLEDGEMENTS

Thank you to my wife Mackenzie and her support throughout my research. Thanks also to my committee members, Dr. Piccolo, Dr. Hansen, and Dr. Ridge, for their guidance and support. I also want to thank Mary-Claire King and Brian Shirts of the University of Washington for the help they provided, as well as the Fulton Supercomputing Lab and the maintainers of The Cancer Genome Atlas. Thanks also to the Simmons Center for Cancer Research for providing funding to support this research.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Approximately 1-5% of breast cancer patients have a germline mutation in either the BRCA1 or BRCA2 gene. These individuals have a 30-70% chance of developing breast cancer in their lifetime. BRCA1 and BRCA2 play important roles in DNA repair, specifically homologous recombination repair (HRR) of double-stranded breaks. When double-stranded breaks have occurred, cells to develop into cancerous tumors rather than enter apoptosis.

Many patients exhibit clinical responses that are characteristic of those who carry germline BRCA1/2 mutations, even though they lack germline BRCA1/2 mutations. This phenomenon, known as "BRCAness", may result from genomic or epigenomic aberrations that have similar, downstream biological effects as germline mutations in BRCA1/2. Such downstream effects may include an increase in double-stranded breaks and other HRR deficiencies, but relatively little is understood about the biological drivers and effects of BRCAness. As a result, relatively little is understood about what specific criteria should be used to assign patients to this category. If reliable BRCAness criteria could be identified, better or more specific treatments for BRCAness patients could be applied. For example, treatments for BRCA1/2 patients commonly include PARP inhibitors and platinum-salt therapies, which target cells with HRR defects. Because there is a possible link between BRCAness and HRR deficiencies, it is possible that these same treatments could be effective for BRCAness patients as well.Various criteria have been proposed to classify patients into the BRCAness category; these criteria include somatic mutations in BRCA1/2, large scale (chromosomal) deletions in HRR genes, tumor-mutational signatures, hypermethylation of BRCA1/2 genes, transcriptional profiles, and germline mutations in HRR genes other than BRCA1 and BRCA2.

For this study, we used a multi-omic approach to further investigate the BRCAness phenomenon and to evaluate criteria used to classify patients into this category. We obtained genomic, epigenomic, and transcriptomic data from The Cancer Genome Atlas (TCGA) for 1054 breast-cancer patients. We hypothesized that breast tumors of patients who meet certain BRCAness criteria would exhibit tumor gene-expression patterns or mutational-signature patterns more similar to patients who carry germline mutations in BRCA1 or BRCA2 than to randomly selected breast cancer patients who do not meet these criteria. Such gene-expression patterns or mutational signatures would suggest that breast-tumor biology is affected similarly by germline BRCA1/2 mutations and these other mechanisms and thus that it may be advisable to treat both groups similarly. By including gene-expression data and mutational-signature data in our analysis, we were able to examine multiple sources of evidence for downstream effects of BRCA1/2 mutations simultaneously. For both data types, we found that there was a statistically significant similarity between patients with germline BRCA1/2 mutations and patients categorized as BRCAness; we did not observe such similarity between BRCAness patients and our control group. Upon examining the criteria we used to classify the patients into the BRCAness category, we observed that similarities in gene expression depended strongly on the categorization criteria being used. DNA hypermethylation status correlated strongly with germline-mutation status; however, the same was not true for CNVs or somatic mutations in BRCA1/2. In addition, we found that tumor-expression patterns from germline BRCA1 patients differed significantly from tumors from patients who carried a germline BRCA2 mutation. In contrast, similarity among these groups was quite consistent when examining mutational-signature profiles. Lastly, we examined germline data for 60 additional breast-cancer predisposition genes and observed high similarity in the mutational-signature data between

tumors from BRCA1/2 carriers and tumors from individuals who carried germline mutations in PALB2, RAD51B, and RAD51C, but not in other genes that play an important role in DNA repair. Our findings suggest that the BRCAness category does have biological and clinical relevance but that care must be taken in deciding which patients to classify under this category.

METHODS

Data preparation

We obtained breast-cancer patient data from TCGA. For each patient, we obtained data on germline mutations, somatic mutations, gene methylation, copy-number variations, and gene-expression levels. Firstly, we used these data to categorize each patient into one of the following categories: BRCA, BRCAness, or Other. Secondly, to determine which criteria could be beneficial in characterizing BRCAness, we used the data to analyze similarities and differences among these groups. Due to the heterogeneous nature of how TCGA data are formatted, it was necessary to reformat the data. Therefore, we wrote computer scripts in the Python programming language (http://python.org) and restructured the data into the "tidy data" format. The *BRCA* group included patients who possessed a germline BRCA1/2 mutation that we deemed to be pathogenic (or likely pathogenic); the *BRCAness* group included patients who lacked a known, pathogenic, germline BRCA1/2 mutation but whose tumor had a somatic mutation (single-nucleotide variant or small insertion/deletion) in BRCA1/2, a homozygous deletion in BRCA1/2, or hypermethylation in BRCA1/2; the *Other* group consisted of patients who were identified as having none of these aberrations.

To determine germline-mutation status, we downloaded raw sequencing data from CGHub for matching normal (blood) samples in TCGA. We limited our analysis to whole-exome sequencing samples that had been sequenced using Illumina Genome Analyzer or HiSeq equipment. Because the sequencing data files were stored in BAM format, we used Picard Tools (*SamToFastq* module, version 1.131) to convert them to FASTQ format. We used the Burrows-Wheeler Alignment (BWA, version 0.7.12) tool to align the sequencing reads to version 19 of the GENCODE reference genome (hg19 compatible). We used *sambamba* (version 0.5.4) to sort, index, mark duplicates, and flag statistics for the aligned BAM files. In cases where there were multiple BAM files per sample, we used *bamUtil* (1.0.13) to merge the BAM files. When searching for relevant germline variants, we focused solely on 62 genes that had been included in the BROCA Cancer Risk Panel (http://tests.labmed.washington.edu/BROCA). We extracted data for these genes using bedtools (*intersectBed* module, version 2).

We used Picard Tools (*CalculateHsMetrics* module) to calculate alignment metrics. For exome-capture regions across all samples, the average sequencing coverage of target regions was 44.4. The average percentage of target bases that achieved at least 30X coverage was 33.7%. The average percentage of target bases that achieved at least 100X coverage was 12.3%.

To call DNA variants, we used *freebayes* (version v0.9.21-18-gc15a283) and *Pindel* (https://github.com/genome/pindel). We used *freebayes* to identify single-nucleotide variants and small insertions or deletions. We used *Pindel* to identify larger variants, including deletions and medium-sized insertions. Having called these variants, we used *snpEff* (version 4.1) to annotate the variants and *GEMINI* (version 0.16.3) to query the variant data. The scripts and code that were used to process the data can be found in this open-access repository: https://bitbucket.org/srp33/tcga_germline/overview. We collaborated with Drs. Mary-Claire King and Brian Shirts from the University of Washington to further filter the germline variants for pathogenicity.

We classified pathogenic, somatic mutations in each patient by examining preprocessed data available from the Genomic Data Commons and using the following exclusion criteria: 1) synonymous variants and variants that *snpEff* classified as having a “LOW” or “MODIFIER” effect on protein sequence, 2) variants that SIFT and Polyphen2 both indicated to be benign, and 3) variants that were observed at greater than 1% frequency across all populations in ExAC

Lastly, we collaborated with our colleagues at the University of Washington to evaluate pathogenicity of the somatic variants in BRCA1/2 and compared these findings against data available in the ClinVar database.

We downloaded DNA methylation data via Synapse (https://www.synapse.org/#!Synapse:syn2320010). These data were generated using the Illumina HumanMethylation450 platform. To map the methylation probes to genes, we used an annotation file (*Closest_TSS_gene_name* column) developed by Price, et al. This file can be accessed from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL16304. In cases where there were multiple values per gene, we used the median value. We classified tumor samples as exhibiting hypermethylation in BRCA1/2 using the *getOutliersI* function in the *extremevalues* R package (version 2.3.2). When invoking this function, we specified the following non-default parameter values:rho=(1, 1) and FLim=(0.1, 0.9).

We obtained copy-number-variation data from the Xena database (https://xenabrowser.net/datapages/?dataset=TCGA.BRCA.sampleMap/Gistic2_CopyNumber_Gistic2_all_thresholded.by_genes&host=https://tcga.xenahubs.net). These data were generated using Affymetrix SNP 6.0 arrays, and CNV calls were made using the GISTIC2 method. The data had been summarized to gene-level values, and CNV values were summarized using integer-based discretization. We focused on tumors with a gene count of “-2” for BRCA1 or BRCA2, which indicated a homozygous deletion in those genes.

We used RNA-Sequencing data that had been preprocessed using the *Rsubread* package and summarized to gene-level values. To facilitate biological and clinical interpretation, we limited the gene-expression data to the The Prosigna™ Breast Cancer Prognostic Gene Signature (PAM50) genes.

We derived a mutational-signature profile for each patient using the *deconstructSigs* (version 1.8.0) R package using the mutational data for each breast-cancer patient. As input to this step, we used somatic-mutation data that had not been filtered for pathogenicity, as a way to ensure adequate representation of each signature. This provided us with a data set that has a value for each of the mutational signatures for each patient, indicating the "weight" of each signature. We formatted mutational data using the *mut.to.sigs.input* function, then used the *whichSignatures* function with the included *signatures.nature2013* data as the signature profile data set to process the data.

Analytical Pipeline

We analyzed the PAM50 gene-expression profiles and mutational-signature profiles for each patient using *Rtsne* (version 0.11), an R package that implements the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm. This algorithm enabled us to further reduce the dimensionality of the data and visualize similarities and differences among tumors based on these gene-expression or mutational-signature patterns.

For a given group or groups of patients, we used Cartesian coordinates produced by the t-SNE algorithm to determine similarity by calculating the pairwise Euclidean distance between each patient in the group(s). We then calculated the median of the pairwise Euclidean distances. To determine whether the similarity within or between groups was statistically significant, we performed a permutation analysis. First, we created an empirical null distribution against which we could compare the actual median distances; to create this distribution, we calculated the median, pairwise Euclidean distance among or between individuals of group(s) the same size after randomizing the "identity" of each sample. We calculated empirical p-values by finding the percentage of randomized medians higher than the actual median.

RESULTS

The purpose of this study was to evaluate similarity between breast-cancer patients with germline BRCA1/2 mutations and BRCAness patients. To do this, we categorized patients into one of three categories (BRCA, BRCAness, Other), and performed an integrative analysis to evaluate gene-expression profiles and mutational-signature profiles of each patient. Figure 1 illustrates patient counts for each of the patient categories. The *Other* group contained the largest number of patients (n = 927), whereas the BRCA group was the smallest, containing only 47 patients.

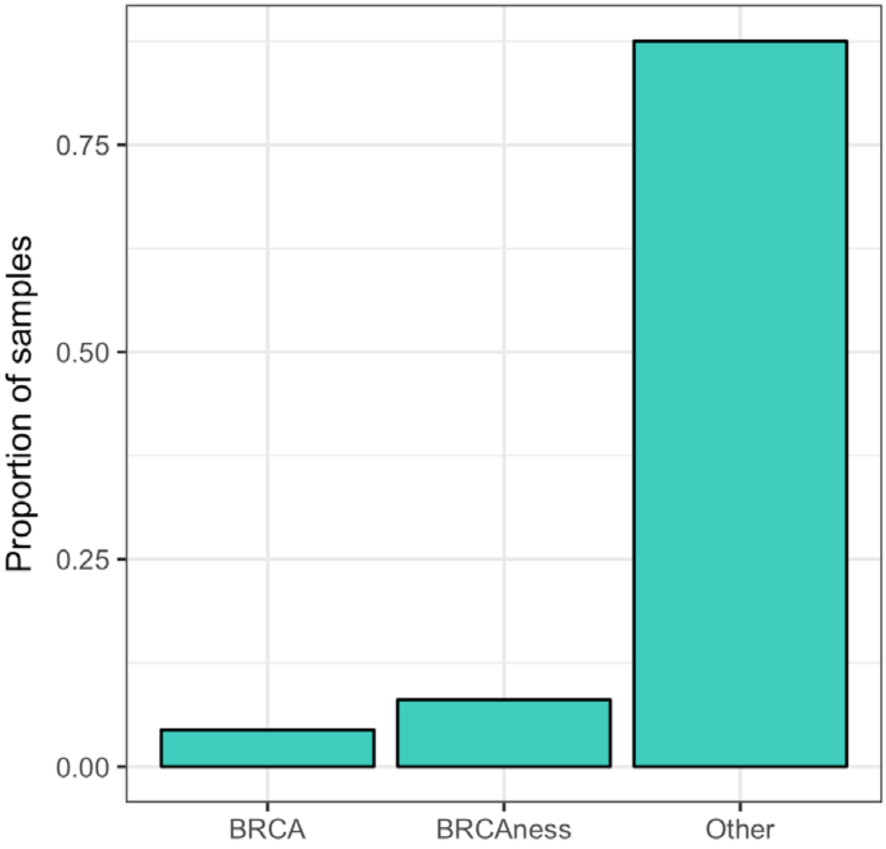

Figure 1 - Distribution of patients in each patient category.

As a way to characterize downstream biological effects that may result from genomic aberrations in BRCA1 or BRCA2, we evaluated gene-expression data from 1054 breast tumors in TCGA. A profile for each patient consisted of expression values for the 50 genes from the PAM50 panel (see Methods), which has been demonstrated to have biological and clinical relevance. We also evaluated mutational signatures for the same cohort of patients (see Methods). These signatures reflect somatic-mutation patterns of single nucleotide variants in a trinucleotide context that likely result from lack of DNA damage repair and other aberrant cellular processes and have been shown to have clinical relevance. This resulted in a data table containing a weight for each signature for each patient. An example of this output is shown in Table 1.

Table 1 - Example output of deconstructSigs.

	TCGA ID	Signature.1A	Signature.1B	Signature.2	Signature.3	...
1	TCGA-B7-XYZ1	0.552	0.000	0.239	0.000	...
2	TCGA-A2-XYZ2	0.000	0.000	1.000	0.000	...
3	TCGA-C3-XYZ3	0.000	0.000	0.000	0.000	...
4	TCGA-D7-XYZ4	0.446	0.000	0.000	0.000	...
5	...	...	...	...	...	...

To reduce the dimensionality of the data, we applied the t-distributed Stochastic Neighbor Embedding (t-SNE) algorithm to the gene-expression and mutational-signature profile data. This algorithm produced X and Y coordinates for each patient, which we then plotted (Figures 2 and 3). These scatterplots illustrate interesting patterns that arise from the data. For the mutational-signature data (Figure 2), samples with germline BRCA1 and BRCA2 mutations cluster primarily in one area that is predominantly populated by “Signature 3” tumors.

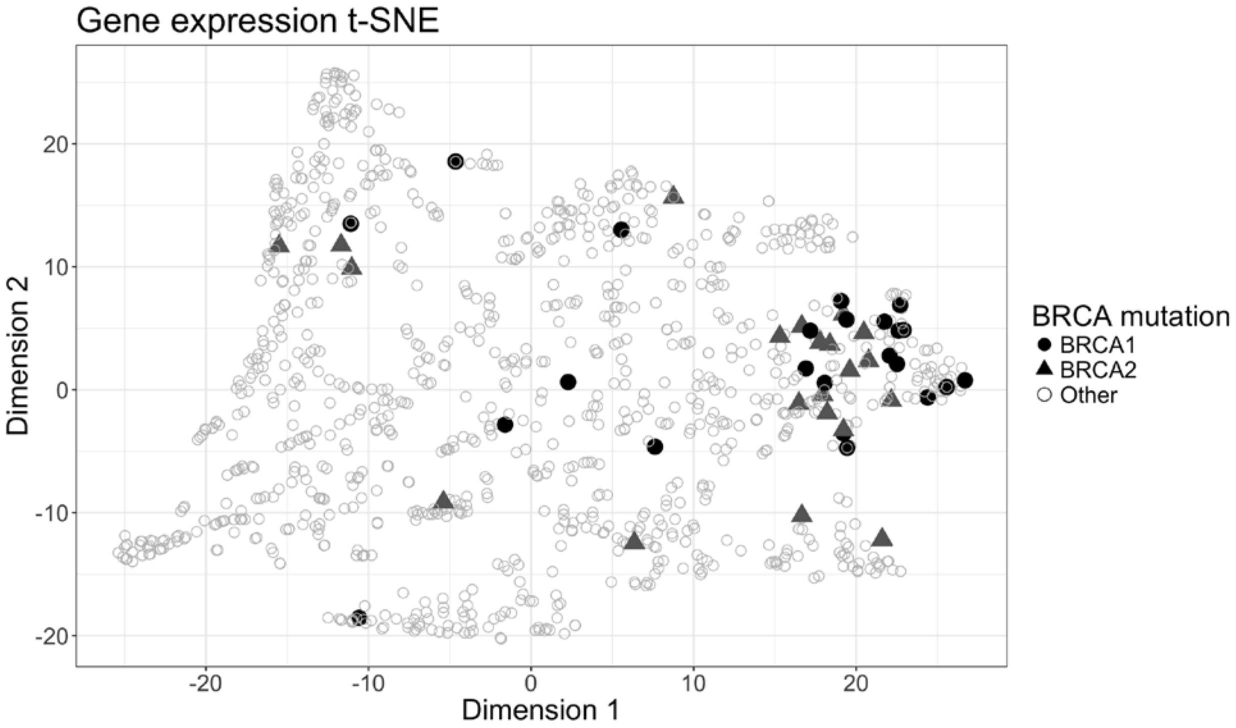

Figure 2 - Scatterplot showing the mutational signature profiles.

The t-SNE plot for the gene-expression data (Figure 3) shows a cluster (upper right) that is distinct from the remaining patients and is populated almost exclusively by tumors of the "Basal" subtype (PAM50 classification). Nearly all of the patients with BRCA1 germline mutations fell in this cluster, whereas only four BRCA2 patients fell in this cluster.

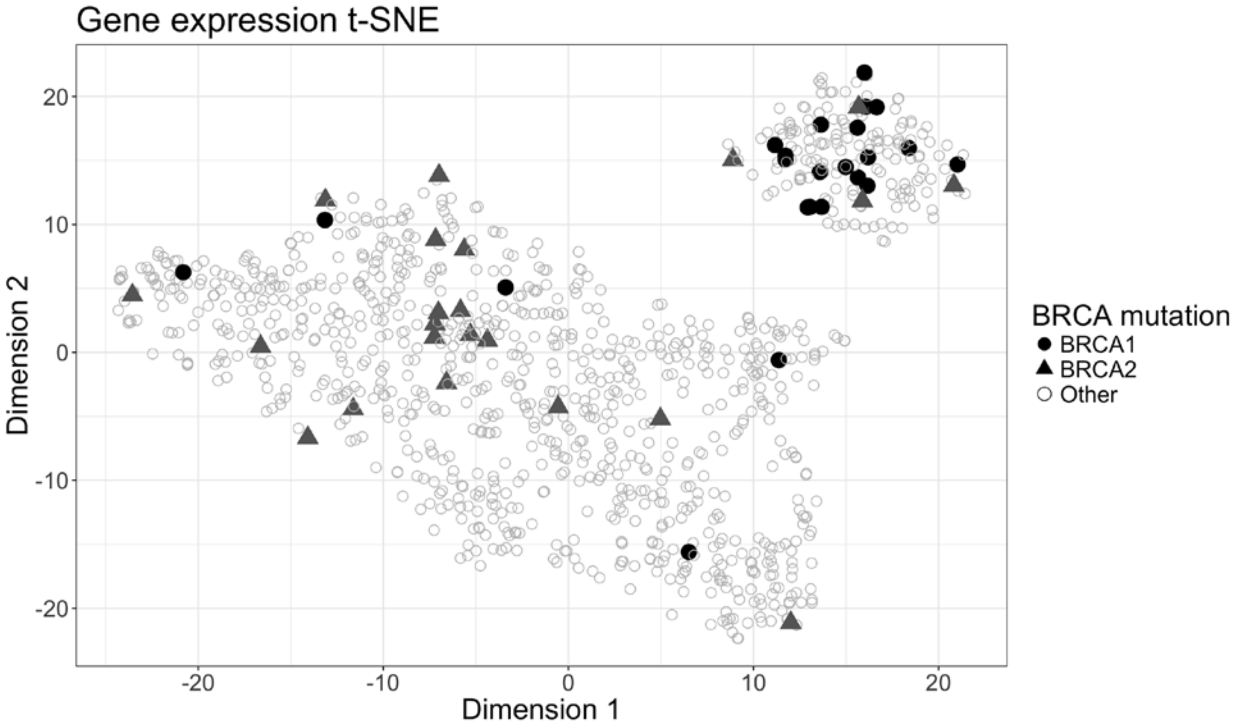

Figure 3 - Scatterplot showing the gene expression profiles.

As a complement to visualizing the data using the t-SNE algorithm, we used a permutation analysis to evaluate the similarity within and between various groups of patients. First, we analyzed the homogeneity among BRCA1 patients when compared to themselves, expecting that there would be a high degree of similarity. The results of this analysis shows significant similarity (P-value < 0.001) within the BRCA1 group, which is expected (Figure 4). However, we do not observe statistically significant similarity within the BRCA2 group in the gene expression context (P-value 0.121). We observe significant similarity within both the BRCA1 group (P-value <0.001) and the BRCA2 (P-value <0.001) group in the mutational signature data (Figure 4).

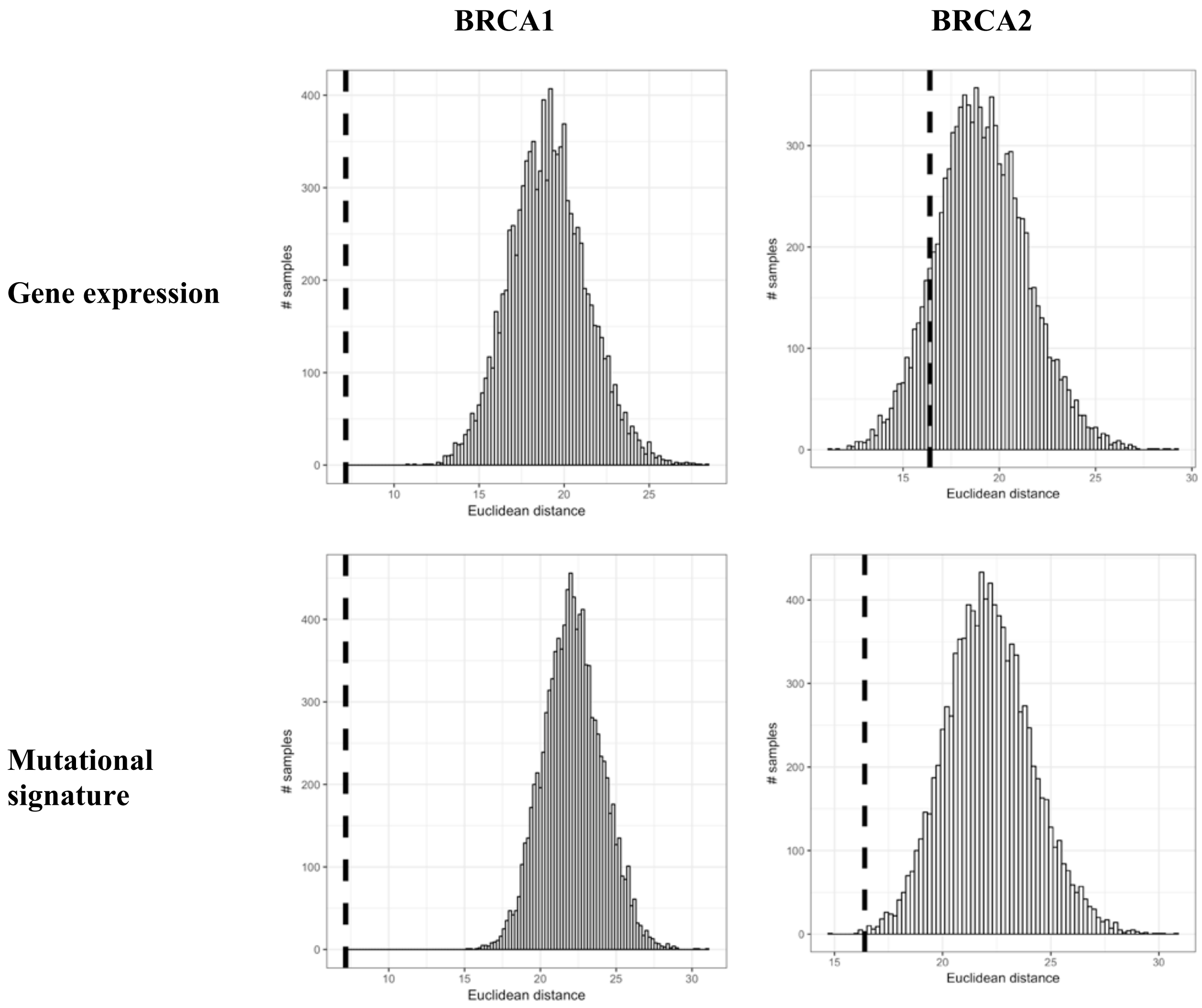

Figure 4 - Permutation analysis comparing BRCA1/2 patients to others with the same mutation across two data sets.

Next we compared patients in the *BRCA* category to those in the *BRCAness* category. As shown in Figure 5, there was a statistically significant similarity between the BRCA and BRCAness groups in the mutational signature data (P-value <0.001). However, when this analysis was repeated with the gene expression data, we found that there isn’t significant similarity between the two groups.

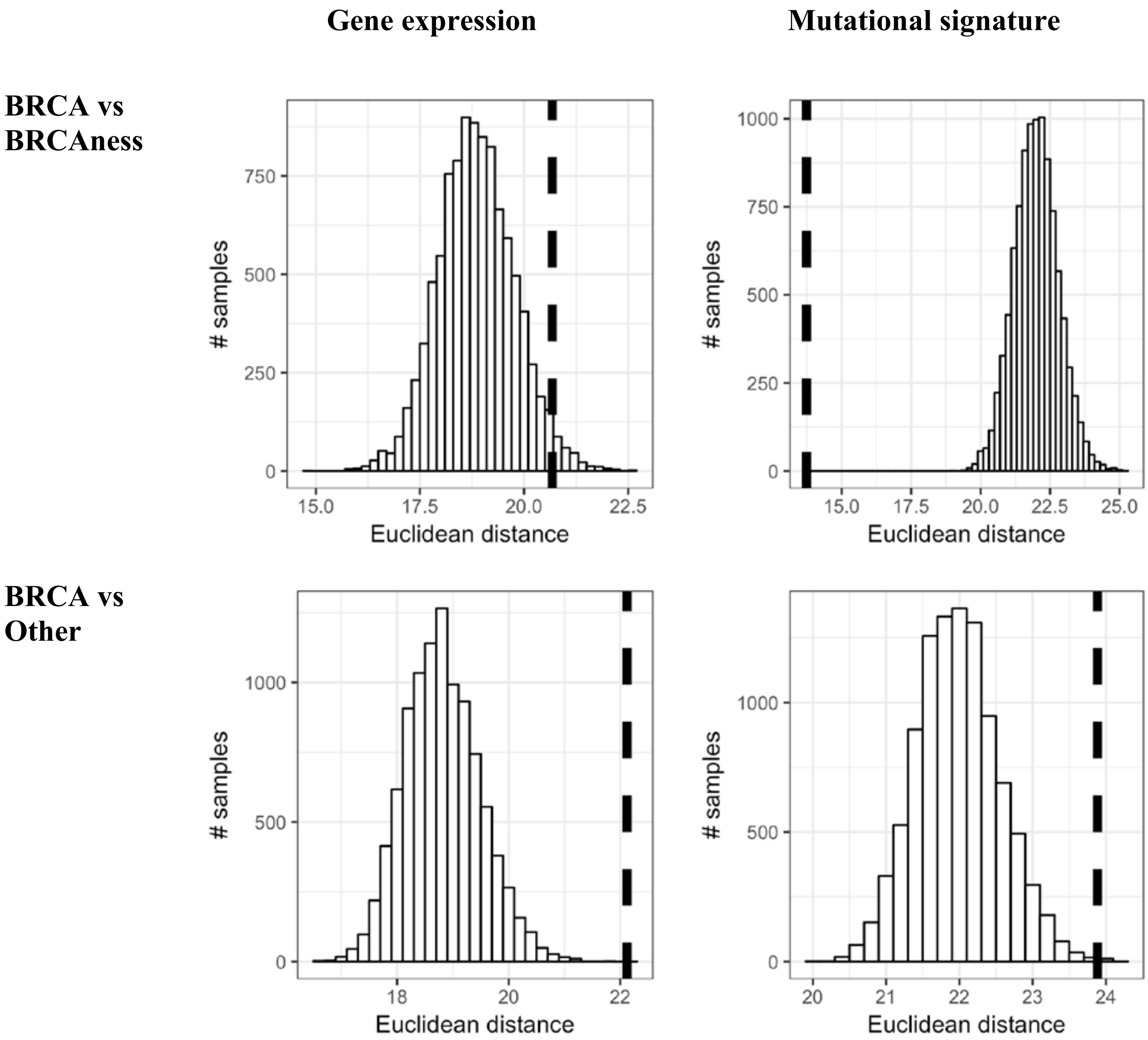

Figure 5 - Results of the permutation analysis of BRCA patients compared to BRCAness patients in two data sets.

A more detailed analysis of the mutational-signature data revealed that all subgroups within the *BRCAness* category—hypermethylation, deletions, or somatic mutations—showed high similarity to patients in the *BRCA* category, irrespective of whether these aberrations affected BRCA1 or BRCA2 (Table 2)

Table 2 - Empirical p-values for subgroup comparisons using the mutational-signature data.

Gene	**Deletion**	**Methylation**	**Somatic mutation**
BRCA 1	<0.001	<0.001	<0.001
BRCA 2	<0.001	0.004	<0.001
BRCA 1&2	<0.001	<0.001	<0.001

A subgroup analysis of the gene-expression data revealed that patients with germline BRCA1 mutations were highly similar to patients with somatic hypermethylation in BRCA1 (Table 3). However, expression patterns for patients with germline BRCA1 mutations were significantly different from patients with somatic BRCA1 mutations. Patients with germline BRCA2 mutations showed similar results. Although only one tumor exhibited hypermethylation in BRCA2, expression patterns for this tumor were significantly similar to patients with somatic BRCA2 hypermethylation. Patients with somatic BRCA2 mutations were significantly dissimilar to patients with patients with germline mutations in this gene.

Table 3 - Empirical p-values for subgroup comparisons using the gene expression data.

Gene	Deletion	Methylation	Somatic mutation
BRCA 1	0.698	<0.001	0.919
BRCA 2	0.921	0.004	0.983
BRCA 1&2	0.821	0.045	0.007

When we considered BRCA1 and BRCA2 separately for the gene-expression, we observed a significant difference between patients with either a BRCA1 germline mutation or tumor hypermethylation and patients with either a BRCA2 germline mutation or somatic hypermethylation (Table 3). However, when we did the same for the mutation signatures, these two subgroups were statistically indistinguishable (Table 2).

In our initial evaluations, we only considered somatic aberrations as candidates for classifying patients into the *BRCAness* category. However, germline mutations in many other genes are known to be breast-cancer predisposition genes and may confer similar downstream effects on tumor biology as BRCA1 or BRCA2. In particular, we were interested in genes that aid in homologous recombination repair. We searched for germline variants in 60 such genes (see Methods). Germline variants occurred most frequently in CHEK2 (n=25) and ATM (n=10) with a long tail of mutations occurring in a variety of other genes (Figure 6).

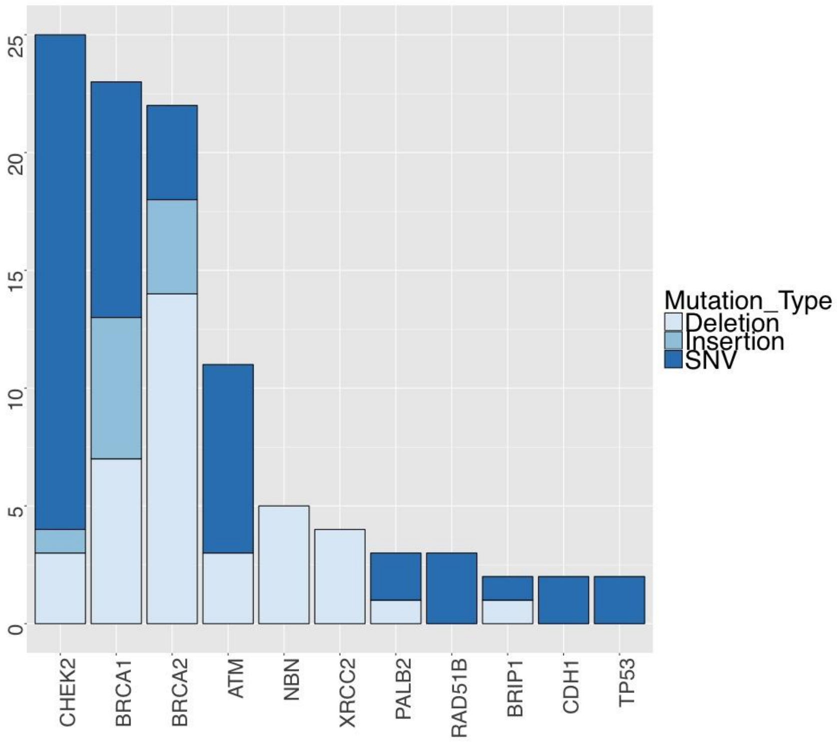

Figure 6 - Distribution of germline mutations across several genes.

We used the gene-expression data and mutational signatures to evaluate these genes as candidates to be included in the BRCAness category. For the gene-expression data, two patients who carried germline mutations in either RAD51B or RAD51C had a tumor of the Basal subtype; however, two additional patients with a mutation in RAD51B did not cluster with these patients. The patterns for the mutational-signature data were more clear. Seven of eight patients who carried a germline mutation in PALB2, BARD1, RAD51B, or RAD51C clustered tightly with the Signature 3 samples, even though Signature 3 was the most prominent signature for only one of these patients. Each of these genes codes for a protein that plays a role in homologous recombination repair and interacts—whether directly or indirectly—with BRCA1 and/or BRCA2. Accordingly, we created a new category called *HRR+* that consisted of patients who had a germline mutation in one of these genes. We then used a permutation analysis to assess the level of similarity between the *HRR+* group and the *BRCA* group. This analysis revealed a highly significant similarity between these groups as well as between the PALB2 mutated samples considered alone (Table 4). However, these relationships were not significant in the gene expression data (Table 4).

Table 4 - P-values of analyses of PALB2 and HRR+ patients.

	BRCA1 Gene expression	BRCA2 Gene expression	BRCA1&2 Gene expression	BRCA1 Mutational signatures	BRCA2 Mutational signatures	BRCA1&2 Mutational signatures
PALB2	0.215	0.231	0.199	<0.001	<0.001	<0.001
HRR+	0.949	0.850	0.923	<0.001	<0.001	<0.001

DISCUSSION

Our analysis comparing mutational signature profiles of *BRCA* patients to *BRCAness* patients revealed statistically significant similarity between the two groups. Additionally, our analysis comparing *BRCA* patients to *Other* patients revealed significant difference between the two groups. The results of this analysis suggest that BRCA patients are more similar to BRCAness patients than they are to Other patients in terms of mutational signature profiles, and also that mutational signature profiles could be an indicator of BRCAness. The results also suggest that in terms of mutational signature data, our method of categorizing BRCAness patients using BRCA1/2 hypermethylation, somatic mutations, and homozygous chromosomal deletions is a valid method for categorizing BRCAness patients. Our results also suggest that additional patients who cluster with the signature 3 patients—especially those who carry germline mutations in PALB2, BARD1, RAD51B, or RAD51C—could be classified into the *BRCAness category*. Future steps could include analyzing if a patient without a germline BRCA1/2 mutation who has a high weight in Signature 3 should be categorized as BRCAness, despite not having the other biomarkers we used to categorize BRCAness.

The analysis between BRCA and BRCAness patients in the gene-expression data did not reveal statistically significant similarity overall, suggesting that BRCAness patients do not necessarily have similar gene expression profiles as BRCA patients under our categorization methods and that the downstream effects of HRR inactivation are less well reflected in gene-expression profiles than they are in mutational signatures. However, comparisons between BRCA and BRCAness patients did reveal similarities between some sub categories of patients for the gene expression data. In particular, there is significant similarity between BRCA patients and patients with hypermethylation in BRCA1 and BRCA2.

When using the mutational-signature data to compare individuals with BRCA1 germline mutations against individuals with BRCA2 mutations, we observed that mutations in these genes have a similar effect on a patient's mutational signature profile. However, there was a significant difference between the gene-expression profiles of patients in these groups. Since the difference between the two groups is significant, it also suggests that germline BRCA1/2 mutations affect a patient’s gene expression profile, but that the effects of each of these genes is different from each other. For medical treatments based off of a patient’s gene expression profile, patients with germline BRCA1/2 mutations should perhaps be considered separately from each other.

In regards to comparisons of patients with germline BRCA1/2 mutations, there is significant similarity in both gene expression profiles and in mutational signature profiles to patients with somatic BRCA1/2 hypermethylation. Also, in the case of the mutational signature profiles, there is also significant similarity between patients with germline BRCA1/2 mutations, and patients with somatic BRCA1/2 large scale deletions, as well as significant similarity between patients with germline BRCA2 mutations, and patients with somatic BRCA2 mutations.

In regards to “BRCAness’, this suggests that somatic hypermethylation in BRCA1/2 is an indicator for BRCAness in breast cancer patients in both mutational signature profiles, and gene expression profiles. This also suggests that large scale BRCA1/2 deletions and somatic mutations could be indicators of BRCAness in regards to mutational signature profiles.

REFERENCES

Adzhubei, I. A., Schmidt, S., Peshkin, L., Ramensky, V. E., Gerasimova, A., Bork, P., … Sunyaev, S. R. (2010, April). A method and server for predicting damaging missense mutations. *Nature Methods*. United States. https://doi.org/10.1038/nmeth0410-248

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A. J. R., Behjati, S., Biankin, A. V, … Stratton, M. R. (2013). Signatures of mutational processes in human cancer. *Nature*, *500*(7463), 415–421. Retrieved from http://dx.doi.org/10.1038/nature12477

Antoniou, A., Pharoah, P. D. P., Narod, S., Risch, H. A., Eyfjord, J. E., Hopper, J. L., … Easton, D. F. (2003). Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies. *American Journal of Human Genetics*, *72*(5), 1117–1130. https://doi.org/10.1086/375033

Beltran, H., Yelensky, R., Frampton, G. M., Park, K., Downing, S. R., MacDonald, T. Y., … Rubin, M. A. (2013). Targeted Next-generation Sequencing of Advanced Prostate Cancer Identifies Potential Therapeutic Targets and Disease Heterogeneity. *European Urology*, *63*(5), 920–926. https://doi.org/http://dx.doi.org/10.1016/j.eururo.2012.08.053

Boland, C. R., & Goel, A. (2010). Microsatellite instability in colorectal cancer. *Gastroenterology*.

Bruin, E. C., McGranahan, N., Mitter, R., Salm, M., Wedge, D. C., & Yates, L. (2014). Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science*.

Ceschin, R., Panigrahy, A., & Gopalakrishnan, V. (2015). sfDM: Open-Source Software for Temporal Analysis and Visualization of Brain Tumor Diffusion MR Using Serial Functional Diffusion Mapping. *Cancer Informatics*, *14*(Suppl 2), 1–9. https://doi.org/10.4137/CIN.S17293

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., … Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Couch, F. J., Johnson, M. R., Rabe, K. G., Brune, K., de Andrade, M., Goggins, M., … Hruban, R. H. (2007). The Prevalence of BRCA2 Mutations in Familial Pancreatic Cancer. *Cancer Epidemiology Biomarkers & Prevention* , *16*(2), 342–346. https://doi.org/10.1158/1055-9965.EPI-06-0783

Dayton, J. B., & Piccolo, S. R. (2017). Classifying cancer genome aberrations by their mutually exclusive effects on transcription. *bioRxiv*. Retrieved from http://biorxiv.org/content/early/2017/03/31/122549.abstract

Dickinson, D. J., Ward, J. D., Reiner, D. J., & Goldstein, B. (2013). Engineering the Caenorhabditis elegans genome using Cas9-triggered homologous recombination. *Nat Meth*, *10*(10), 1028–1034. Retrieved from http://dx.doi.org/10.1038/nmeth.2641

Evan, G. I., & Vousden, K. H. (2001). Proliferation, cell cycle and apoptosis in cancer. *Nature*, *411*(6835), 342–348. Retrieved from http://dx.doi.org/10.1038/35077213

Gallagher, D. J., Konner, J. A., Bell-McGuinn, K. M., Bhatia, J., Sabbatini, P., Aghajanian, C. A., … Kauff, N. D. (2011). Survival in epithelial ovarian cancer: a multivariate analysis incorporating BRCA mutation status and platinum sensitivity. *Annals of Oncology* , *22*(5), 1127–1132. https://doi.org/10.1093/annonc/mdq577

Goldman, M., Craft, B., Swatloski, T., Cline, M., Morozova, O., Diekhans, M., … Zhu, J. (2015). The UCSC Cancer Genomics Browser: update 2015. *Nucleic Acids Research*, *43*(D1), D812–D817. Retrieved from http://dx.doi.org/10.1093/nar/gku1073

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., … Lander, E. S. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science (New York, N.Y.)*, *286*(5439), 531–537.

Grossman, R. L., Heath, A. P., Ferretti, V., Varmus, H. E., Lowy, D. R., Kibbe, W. A., & Staudt, L. M. (2016). Toward a Shared Vision for Cancer Genomic Data. *New England Journal of Medicine*, *375*(12), 1109–1112. https://doi.org/10.1056/NEJMp1607591

Gutman, D. A., Cooper, L. A. D., Hwang, S. N., Holder, C. A., Gao, J., Aurora, T. D., … Brat, D. J. (2013). MR Imaging Predictors of Molecular Profile and Survival: Multi-institutional Study of the TCGA Glioblastoma Data Set. *Radiology*, *267*(2), 560–569. https://doi.org/10.1148/radiol.13120118

Hall, J. M., Lee, M. K., Newman, B., Morrow, J. E., Anderson, L. A., Huey, B., & King, M. C. (1990). Linkage of early-onset familial breast cancer to chromosome 17q21. *Science*, *250*(4988), 1684–1689. Retrieved from http://science.sciencemag.org/content/250/4988/1684.abstract

Harrow, J., Frankish, A., Gonzalez, J. M., Tapanari, E., Diekhans, M., Kokocinski, F., … Hubbard, T. J. (2012). GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Research*, *22*(9), 1760–1774. https://doi.org/10.1101/gr.135350.111

Helleday, T., Eshtad, S., & Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet.*, *15*. https://doi.org/10.1038/nrg3729

John, E. M., Miron, A., Gong, G., Phipps, A. I., Felberg, A., Li, F. P., … Whittemore, A. S. (2007). Prevalence of pathogenic BRCA1 mutation carriers in 5 US racial/ethnic groups. *JAMA*, *298*(24), 2869–2876. https://doi.org/10.1001/jama.298.24.2869

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protocols*, *4*(8), 1073–1081. Retrieved from http://dx.doi.org/10.1038/nprot.2009.86

Landrum, M. J., Lee, J. M., Riley, G. R., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Research*, *42*(D1), D980–D985. Retrieved from http://dx.doi.org/10.1093/nar/gkt1113

Ledermann, J., Harter, P., Gourley, C., Friedlander, M., Vergote, I., Rustin, G., … Matulonis, U. (2012). Olaparib Maintenance Therapy in Platinum-Sensitive Relapsed Ovarian Cancer. *New England Journal of Medicine*, *366*(15), 1382–1392. https://doi.org/10.1056/NEJMoa1105535

Lek, M., Karczewski, K. J., Minikel, E. V, Samocha, K. E., Banks, E., Fennell, T., … Consortium, E. A. (2016). Analysis of protein-coding genetic variation in 60,706 humans. *Nature*, *536*(7616), 285–291. Retrieved from http://dx.doi.org/10.1038/nature19057

Li, H., & Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England)*, *25*(14), 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Liao, Y., Smyth, G. K., & Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, *30*(7), 923–930. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btt656

Lord, C. J., & Ashworth, A. (2016). BRCAness revisited. *Nat Rev Cancer*, *16*(2), 110–120. Retrieved from http://dx.doi.org/10.1038/nrc.2015.21

Ma, C. X., Reinert, T., Chmielewska, I., & Ellis, M. J. (2015). Mechanisms of aromatase inhibitor resistance. *Nat Rev Cancer*, *15*(5), 261–275. Retrieved from http://dx.doi.org/10.1038/nrc3920

MacNeil, S. M., Johnson, W. E., Li, D. Y., Piccolo, S. R., & Bild, A. H. (2015). Inferring pathway dysregulation in cancers from multiple types of omic data. *Genome Medicine*, *7*(1), 61. https://doi.org/10.1186/s13073-015-0189-4

Malone, K. E., Daling, J. R., Doody, D. R., Hsu, L., Bernstein, L., Coates, R. J., … Ostrander, E. A. (2006). Prevalence and predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in white and black American women ages 35 to 64 years. *Cancer Research*, *66*(16), 8297–8308. https://doi.org/10.1158/0008-5472.CAN-06-0503

McGranahan, N., Favero, F., Bruin, E. C., Juul Birkbak, N., Szallasi, Z., & Swanton, C. (2015). Clonal status of actionable driver events and the timing of mutational processes in cancer evolution. *Sci Transl Med*.

Mermel, C. H., Schumacher, S. E., Hill, B., Meyerson, M. L., Beroukhim, R., & Getz, G. (2011). GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biology*, *12*(4), R41. https://doi.org/10.1186/gb-2011-12-4-r41

Murugaesu, N., Wilson, G., Birkbak, N. J., Watkins, T. B., McGranahan, N., & Kumar, S. (2015). Tracking the genomic evolution of esophageal adenocarcinoma through neoadjuvant chemotherapy. *Cancer Discovery*.

Nielsen, T. O., Parker, J. S., Leung, S., Voduc, D., Ebbert, M., Vickery, T., … Ellis, M. J. (2010). A Comparison of PAM50 Intrinsic Subtyping with Immunohistochemistry and Clinical Prognostic Factors in Tamoxifen-Treated Estrogen Receptor–Positive Breast Cancer. *Clinical Cancer Research*, *16*(21), 5222 LP-5232. Retrieved from http://clincancerres.aacrjournals.org/content/16/21/5222.abstract
Paila U, Chapman BA, Kirchner R, Q. A. (2013). GEMINI: Integrative Exploration of Genetic Variation and Genome Annotations. *PLoS Comput Biol*, *9*(7). Retrieved from https://doi.org/10.1371/journal.pcbi

Peng, G., Chun-Jen Lin, C., Mo, W., Dai, H., Park, Y.-Y., Kim, S. M., … Lin, S.-Y. (2014). Genome-wide transcriptome profiling of homologous recombination DNA repair. *Nat Commun*, *5*. Retrieved from http://dx.doi.org/10.1038/ncomms4361

Pfeifer, G. P. (2010). Environmental exposures and mutational patterns of cancer genomes. *Genome Med.*, *2*. https://doi.org/10.1186/gm175

Piccolo, S. R., & Frey, L. J. (2013). Clinical and molecular models of glioblastoma multiforme survival. *International Journal of Data Mining and Bioinformatics*, *7*(3), 245–265.

Powell, S. N., & Kachnic, L. A. (n.d.). Roles of BRCA1 and BRCA2 in homologous recombination, DNA replication fidelity and the cellular response to ionizing radiation. *Oncogene*, *22*(37), 5784–5791. Retrieved from http://dx.doi.org/10.1038/sj.onc.1206678

Price, M. E., Cotton, A. M., Lam, L. L., Farre, P., Emberly, E., Brown, C. J., … Kobor, M. S. (2013). Additional annotation enhances potential for biologically-relevant analysis of the Illumina Infinium HumanMethylation450 BeadChip array. *Epigenetics & Chromatin*, *6*(1), 4. https://doi.org/10.1186/1756-8935-6-4

Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., & Piccolo, S. R. (2015). Alternative preprocessing of RNA-Sequencing data in The Cancer Genome Atlas leads to improved analysis results. *Bioinformatics*, *31*(22), 3666–3672. Retrieved from http://dx.doi.org/10.1093/bioinformatics/btv377

Roberts, N. J., Jiao, Y., Yu, J., Kopelovich, L., Petersen, G. M., Bondy, M. L., … Klein, A. P. (2012). ATM Mutations in Patients with Hereditary Pancreatic Cancer. *Cancer Discovery* , *2*(1), 41–46. https://doi.org/10.1158/2159-8290.CD-11-0194

Robinson, D., Van Allen, E. M., Wu, Y.-M., Schultz, N., Lonigro, R. J., Mosquera, J.-M., … Chinnaiyan, A. M. (2015). Integrative Clinical Genomics of Advanced Prostate Cancer. *Cell*, *161*(5), 1215–1228. https://doi.org/http://dx.doi.org/10.1016/j.cell.2015.05.001

Rosenthal, R., McGranahan, N., Herrero, J., Taylor, B. S., & Swanton, C. (2016). deconstructSigs: delineating mutational processes in single tumors distinguishes DNA repair deficiencies and patterns of carcinoma evolution. *Genome Biology*, *17*(1), 31. https://doi.org/10.1186/s13059-016-0893-4

Shirts, B. H., Casadei, S., Jacobson, A. L., Lee, M. K., Gulsuner, S., Bennett, R. L., … Pritchard, C. C. (2016). Improving performance of multigene panels for genomic analysis of cancer predisposition. *Genetics in Medicine : Official Journal of the American College of Medical Genetics*, *18*(10), 974–981. https://doi.org/10.1038/gim.2015.212

Stratton, M. R., & Rahman, N. (2008). The emerging landscape of breast cancer susceptibility. *Nat Genet*, *40*(1), 17–22. Retrieved from http://dx.doi.org/10.1038/ng.2007.53

Suehiro, Y., Okada, T., Shikamoto, N., Zhan, Y., Sakai, K., Okayama, N., … Sasaki, K. (2013). Germline copy number variations associated with breast cancer susceptibility in a Japanese population. *Tumour Biology*, *34*(2), 947–952. https://doi.org/10.1007/s13277-012-0630-x

Tarasov, A., Vilella, A. J., Cuppen, E., Nijman, I. J., & Prins, P. (2015). Sambamba: fast processing of NGS alignment formats. *Bioinformatics (Oxford, England)*, *31*(12), 2032–2034. https://doi.org/10.1093/bioinformatics/btv098

Turner, N., Tutt, A., & Ashworth, A. (2004). Hallmarks of “BRCAness” in sporadic cancers. *Nat Rev Cancer*, *4*(10), 814–819. Retrieved from http://dx.doi.org/10.1038/nrc1457

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2002). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi1110s43

Vettore, A. L., Ramnarayanan, K., Poore, G., Lim, K., Ong, C. K., Huang, K. K., … Iyer, N. G. (2015). Mutational landscapes of tongue carcinoma reveal recurrent mutations in genes of therapeutic and prognostic relevance. *Genome Medicine*, *7*(1), 98. https://doi.org/10.1186/s13073-015-0219-2

Walsh, T., Lee, M. K., Casadei, S., Thornton, A. M., Stray, S. M., Pennil, C., … King, M.-C. (2010). Detection of inherited mutations for breast and ovarian cancer using genomic capture and massively parallel sequencing. *Proceedings of the National Academy of Sciences* , *107*(28), 12629–12633. https://doi.org/10.1073/pnas.1007983107

Wickham, H. (2007). Reshaping data with the reshape Package. *J Stat Softw*, *21*. https://doi.org/10.18637/jss.v021.i12
Wickham, H. (2014). Tidy Data. *Journal of Statistical Software; Vol 1, Issue 10 (2014)* . https://doi.org/10.18637/jss.v059.i10

Wilks, C., Cline, M. S., Weiler, E., Diehkans, M., Craft, B., Martin, C., … Maltbie, D. (2014). The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database : The Journal of Biological Databases and Curation*, *2014*. https://doi.org/10.1093/database/bau093

Alexandrov LB, Nik-Zainal S, Wedge DC, Campbell PJ, Stratton MR. Deciphering signatures of mutational processes operative in human cancer. Cell. 2013. http://dx.doi.org/10.1016/j.celrep.2012.12.008. (n.d.). Retrieved from http://dx.doi.org/10.1016/j.celrep.2012.12.008

Pages H. BSgenome: Infrastructure for Biostrings-based genome data packages. R package version 1.36.0. (n.d.).

Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group. (2000). *British Journal of Cancer*, *83*(10), 1301–1308. https://doi.org/10.1054/bjoc.2000.1407

APPENDIX: SUPPLEMENTARY MATERIALS

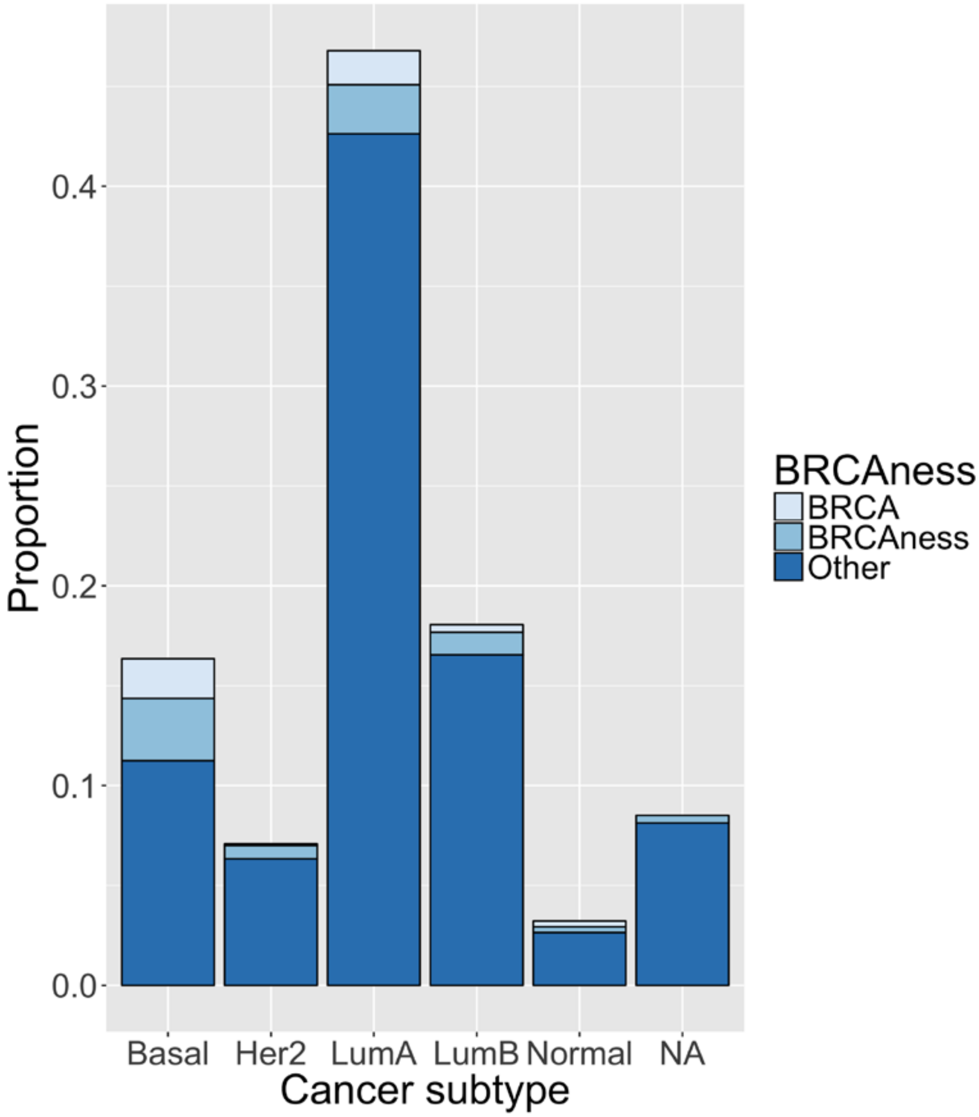

Figure A.1 - Bar chart showing the proportion of patients in each subtype, colored by what BRCA category they fall in.

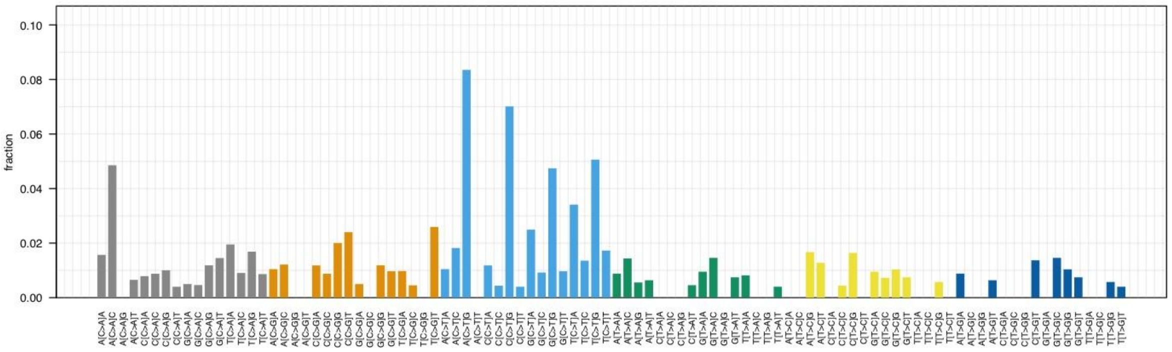

Figure A.2 - Example of one patient's mutational signature.

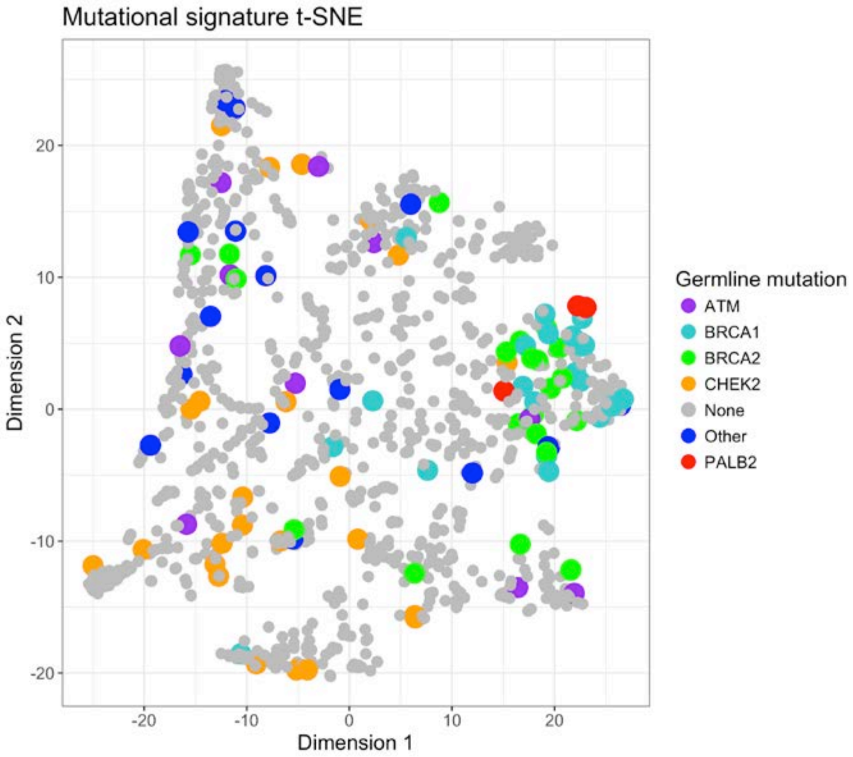

Figure A.3 - Scatterplot showing selected germline mutations in samples in the mutational signature data.

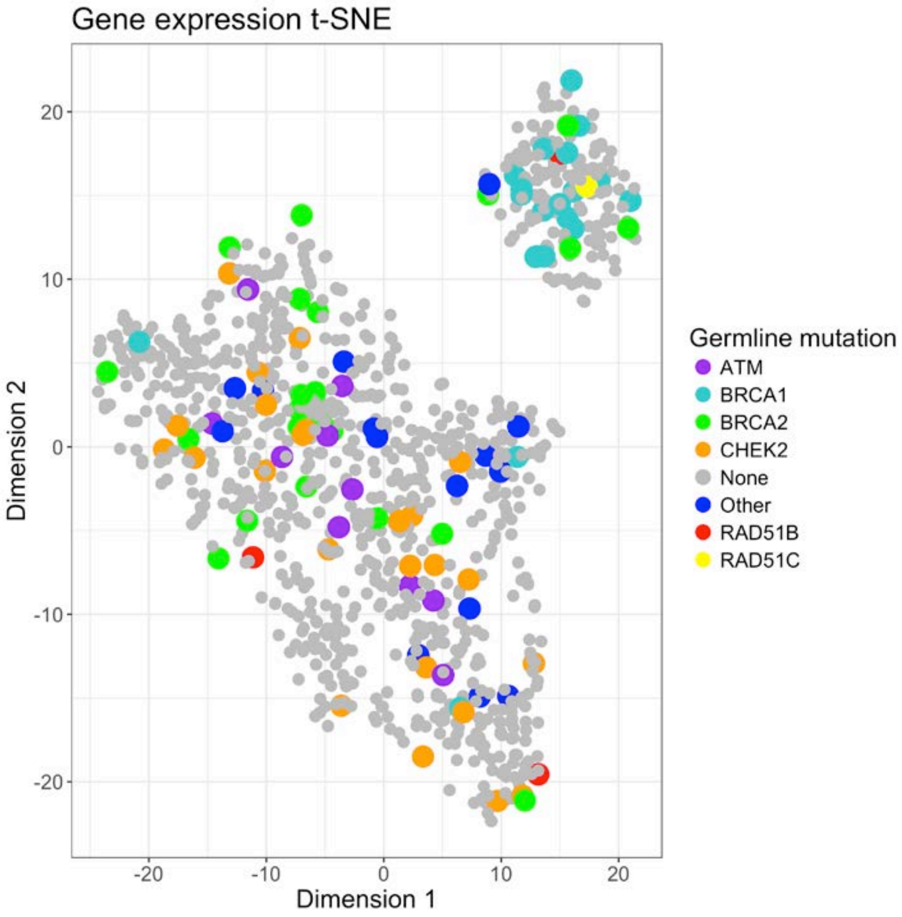

Figure A.4 - Scatterplot showing selected germline mutations in samples in the gene expression data.