NAME: Jitesh Patel
Patel ID: 50079988

Predicting breast cancer survival based on gene expression and clinical variables

Literature Review and Exploratory Analysis

Cancer is a complex disease which involves number of epigenetic and genetic irregularities. Tumors originating in the same tissue or organ may vary considerably in genomic alteration and similar patterns of genomic alteration are observed in tumors from different tissues of origin. Different set of genes are expressed at different stages, types. These gives an opportunity to analyze combination of gene expression and clinical features such as number of lymph nodes, stage types, age etc.

Cancer therapy is challenged by the diversity of molecular implementations of oncogenic process and by the resulting variations in therapeutic responses (Giovanni et al., 2013). Targetable functional events in a tumor class are suggestive of class-specific combination therapy. These may assist in the definition of clinical trials to match actionable oncogenic signatures with personalized therapies (Giovanni et al., 2013). Combining the predictive strength of multiple gene signatures improves prediction of breast cancer survival (Xi Zhao et al., 2011). Mutations in transcriptional factors/regulators show tissue specificity, whereas histone modifiers are often mutated across several cancer types. Clinical association analysis identifies genes having a significant effect on survival, and investigations of mutation with respect to clonal/sub-clonal architecture delineate their temporal orders during tumorigenesis. Taken together, these results lay the groundwork for developing new diagnostics and individualizing cancer treatment (Cyriac Kandoth et al., 2013). Different therapeutic treatment cab more effective for individual patient can be based on combination of gene expression at different cancer stages or age category. Project such as The Cancer Genome Atlas (TCGA) provide molecular tumor maps in unprecedented detail (Giovanni et al., 2013).

I have selected set of genes which are directly or indirectly associated with breast cancer disease from wide range for the analysis. Selected genes are as below.

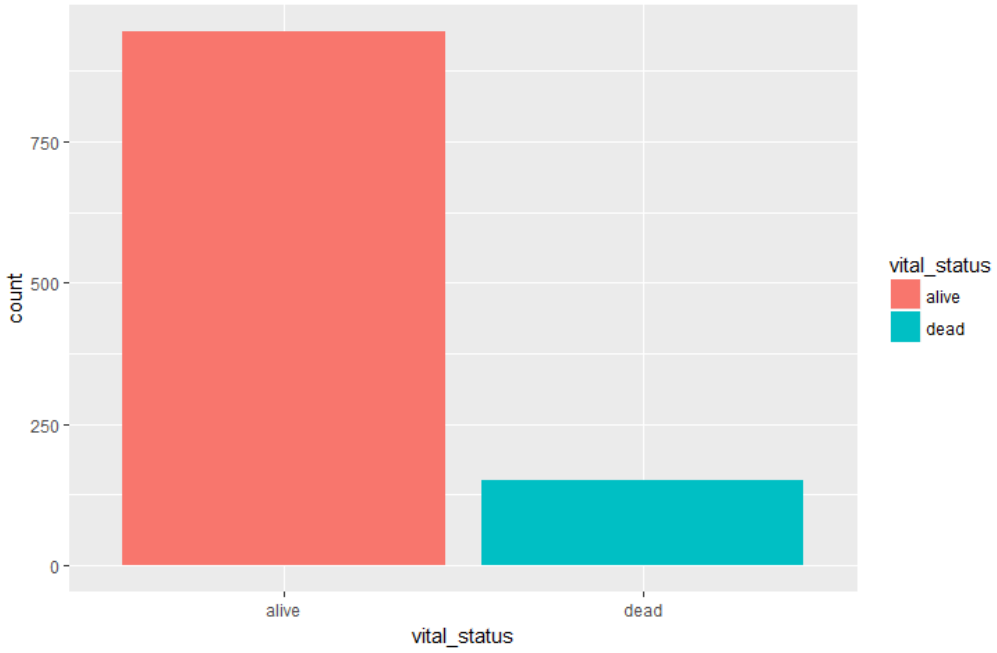
Gene	GeneCards Summary
TP53	TP53 (Tumor Protein P53) is a Protein Coding gene. Diseases associated with TP53 include Li-Fraumeni Syndrome and Choroid Plexus Papilloma. Among its related pathways are Glioma and HTLV-I infection. GO annotations related to this gene include transcription factor activity, sequence-specific DNA binding and protein heterodimerization activity. An important paralog of this gene is TP73.
STK11	STK11 (Serine/Threonine Kinase 11) is a Protein Coding gene. Diseases associated with STK11 include Peutz-Jeghers Syndrome and Testicular Germ Cell Tumor. Among its related pathways are Integrated Breast Cancer Pathway and Metabolism. GO annotations related to this gene include transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity. An important paralog of this gene is CAMKK2.
RB1	RB1 (RB Transcriptional Corepressor 1) is a Protein Coding gene. Diseases associated with RB1 include Retinoblastoma and Small Cell Cancer Of The Lung, Somatic. Among its related pathways are Regulation of retinoblastoma protein and Regulation of activated PAK-2p34 by proteasome mediated degradation. GO annotations related to this gene include transcription factor activity, sequence-specific DNA binding and enzyme binding. An important paralog of this gene is RBL2.

PRKAA2	PRKAA2 (Protein Kinase AMP-Activated Catalytic Subunit Alpha 2) is a Protein Coding gene. Diseases associated with PRKAA2 include Wolff-Parkinson-White Syndrome and Peutz-Jeghers Syndrome. Among its related pathways are Metabolism and mTOR signaling pathway (KEGG). GO annotations related to this gene include transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity. An important paralog of this gene is PRKAA1.
AKT1	AKT1 (AKT Serine/Threonine Kinase 1) is a Protein Coding gene. Diseases associated with AKT1 include Proteus Syndrome, Somatic and Cowden Syndrome 6. Among its related pathways are VEGF Signaling Pathway and Signaling by GPCR. GO annotations related to this gene include identical protein binding and protein kinase activity. An important paralog of this gene is AKT3.
AKT2	AKT2 (AKT Serine/Threonine Kinase 2) is a Protein Coding gene. Diseases associated with AKT2 include Hypoinsulinemic Hypoglycemia With Hemihypertrophy and Diabetes Mellitus, Noninsulin-Dependent. Among its related pathways are VEGF Signaling Pathway and Signaling by GPCR. GO annotations related to this gene include transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity. An important paralog of this gene is AKT1.
AKT3	AKT3 (AKT Serine/Threonine Kinase 3) is a Protein Coding gene. Diseases associated with AKT3 include Megalencephaly-Polymicrogyria-Polydactyly-Hydrocephalus Syndrome 2 and Mpph Syndrome. Among its related pathways are VEGF Signaling Pathway and Signaling by GPCR. GO annotations related to this gene include transferase activity, transferring phosphorus-containing groups and protein tyrosine kinase activity. An important paralog of this gene is AKT1.
MYC	MYC (V-Myc Avian Myelocytomatosis Viral Oncogene Homolog) is a Protein Coding gene. Diseases associated with MYC include Burkitt Lymphoma and Leukemia, Acute Lymphoblastic 3. Among its related pathways are Regulation of nuclear SMAD2/3 signaling and HTLV-I infection. GO annotations related to this gene include transcription factor activity, sequence-specific DNA binding and RNA polymerase II core promoter proximal region sequence-specific DNA binding. An important paralog of this gene is MYCN.
MYCL	MYCL (V-Myc Avian Myelocytomatosis Viral Oncogene Lung Carcinoma Derived Homolog) is a Protein Coding gene. Diseases associated with MYCL include Apocrine Adenosis Of Breast and Lower Lip Cancer. GO annotations related to this gene include transcription factor activity, sequence-specific DNA binding and protein dimerization activity. An important paralog of this gene is MYCN.
MYCN	MYCN (V-Myc Avian Myelocytomatosis Viral Oncogene Neuroblastoma Derived Homolog) is a Protein Coding gene. Diseases associated with MYCN include Feingold Syndrome and Neuroblastoma. Among its related pathways are Transcriptional misregulation in cancer and Neuroscience. GO annotations related to this gene include transcription factor activity, sequence-specific DNA binding and protein dimerization activity. An important paralog of this gene is MYC.

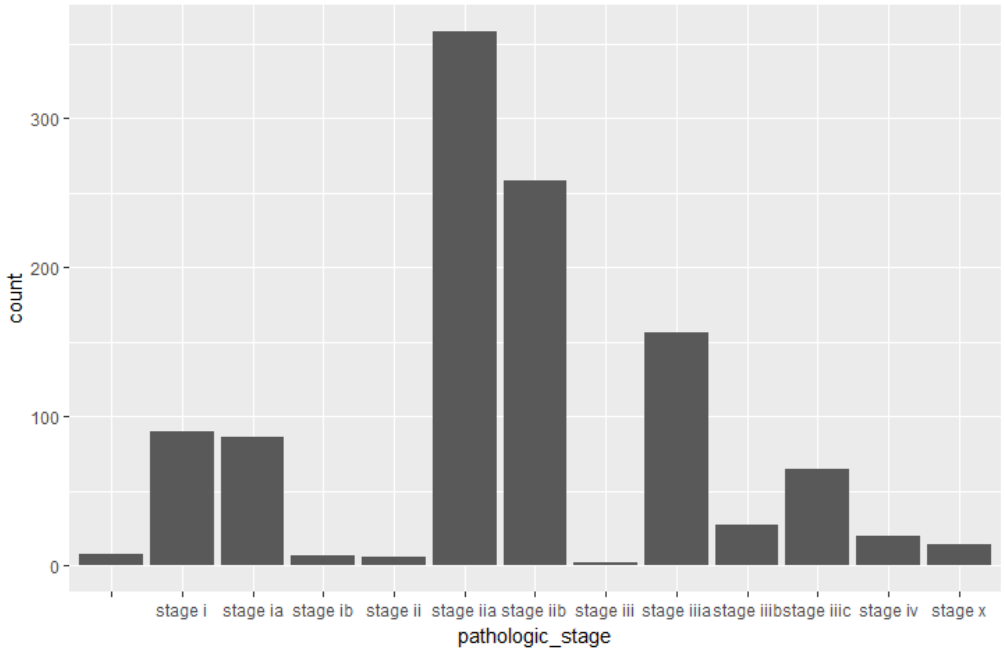
I have downloaded clinical data for breast cancer from TCGA using FireBrowse web API for R. The dataset contains 1097 records with 111 columns for different clinical features. I am going to use some of the clinical variables as below.

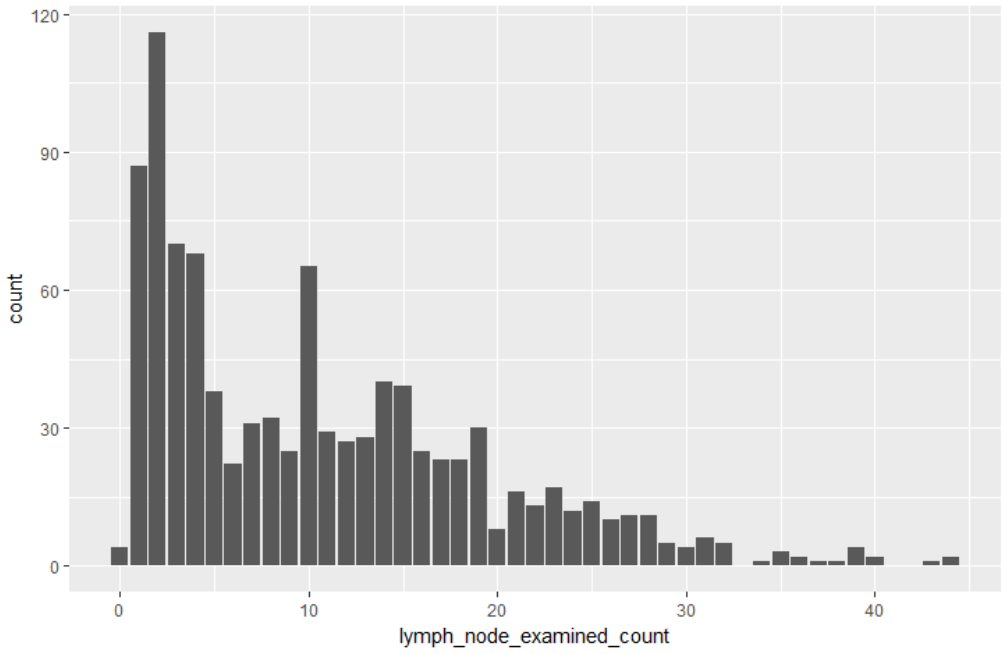
vital_status	Death or alive
pathologic_stage	Cancer stage
lymph_node_examined_count	Number of lymphnodes
gender	Gender
days_to_death	Number of days to death
days_to_last_followup	Number of days from last followup
days_to_birth	Number of days to birth
age_at_initial_pathologic_diagnosis	Age in YEAR at initial diagnosis

Distributions of these variables are as below. Maximum number of patients were alive at the time when I have downloaded the dataset. Very less records with dead as viatal_status.

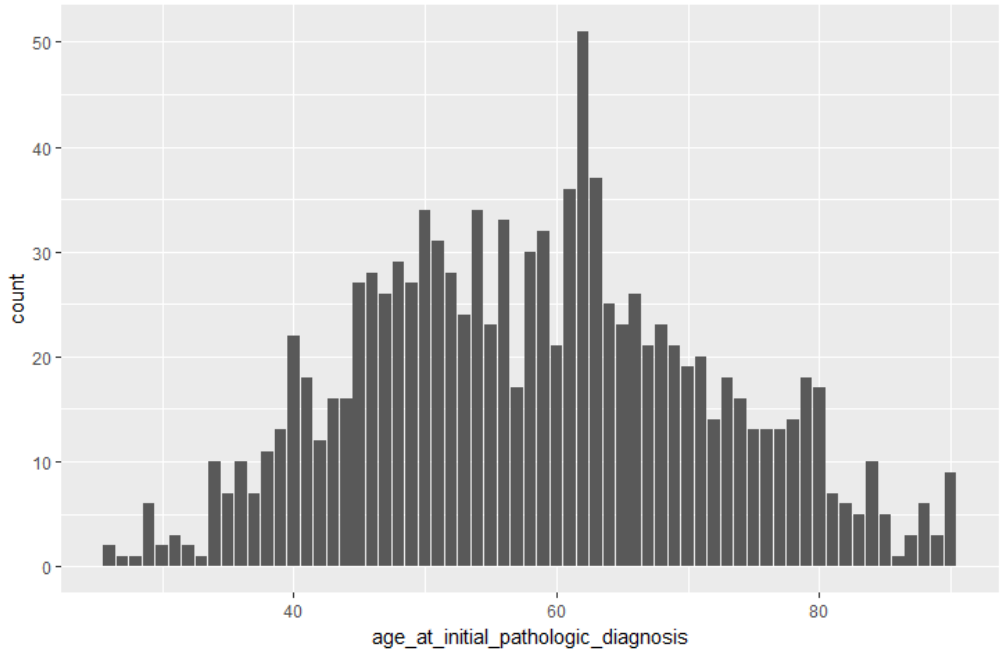


Patients with Stage iia and stage iib are highest. This variable has some of the data. 8 records don't have stage details in the dataset.

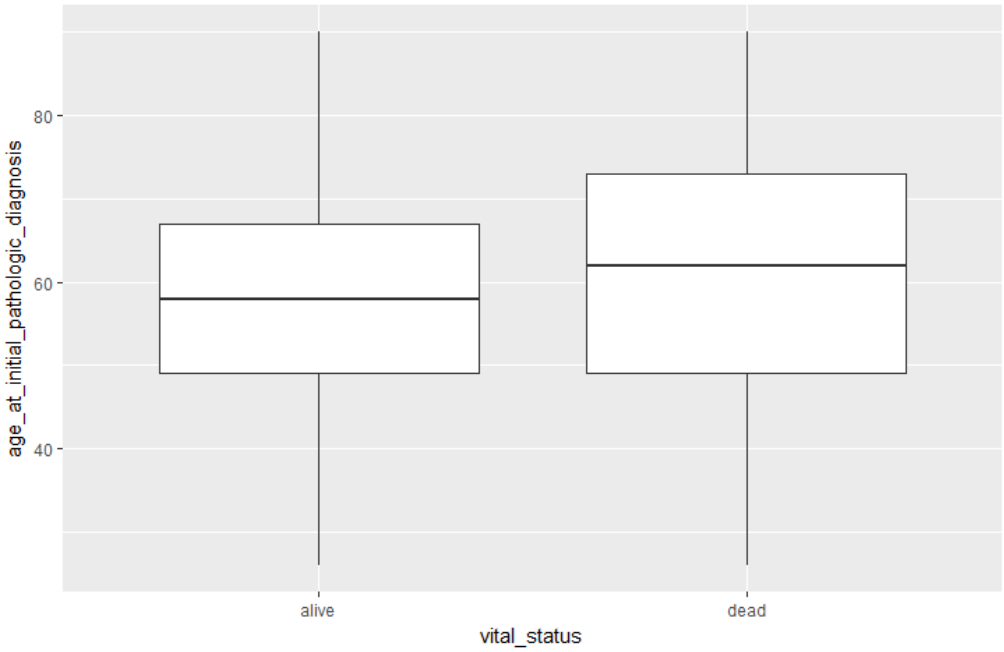




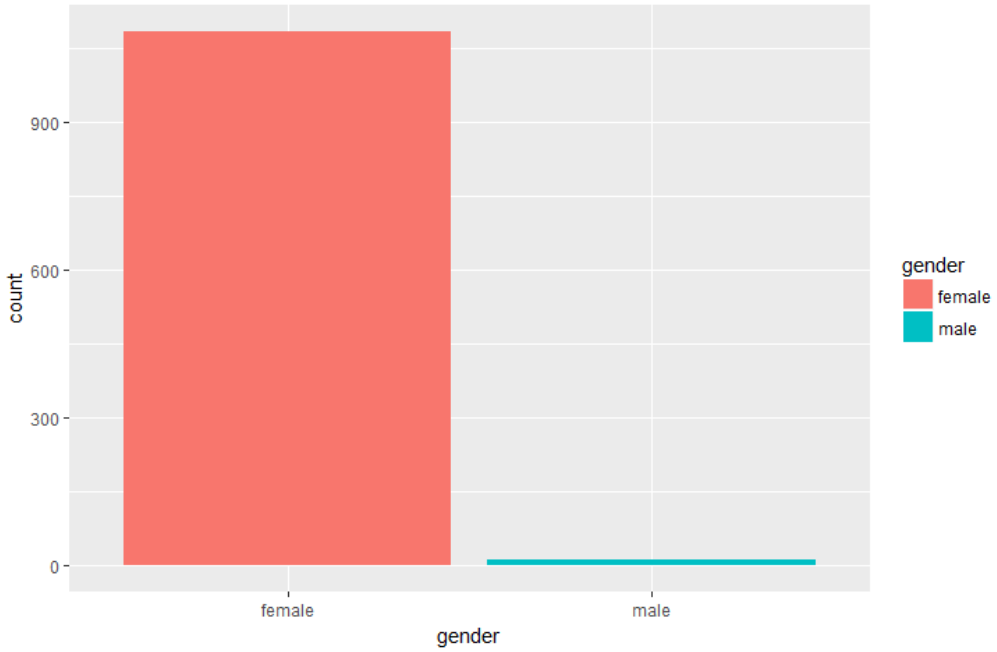
Records with 2 examined lymphnodes are highest and 126 records don’t have details about this variable.



Maximum patients were at age of 62 at the time of initial pathologic diagnosis. Less cases with diagnosis at early stage of the life.

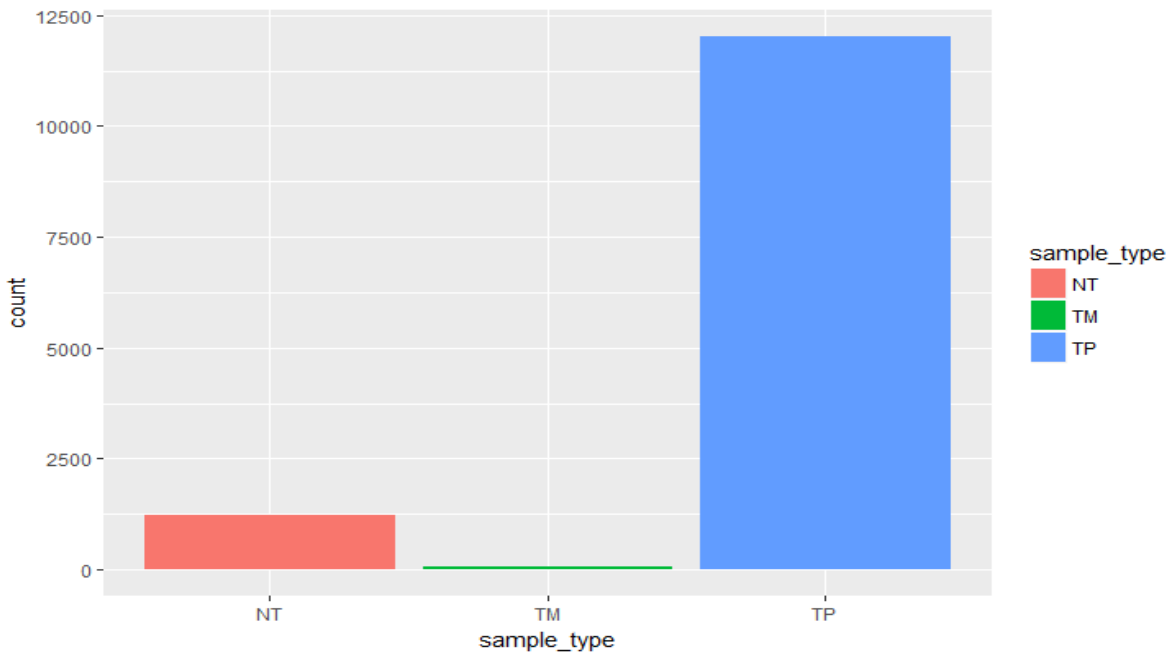


Records with dead as vital_status are more variable than alive. Median for dead is little bit higher than the median of alive.



There are 12 male records with breast cancer and 1085 female records.

To download mRNA expression and Z.SCORE for my selected gene of interest, I have used Barcode present in the clinical dataset identical for each record and downloaded expression of selected genes using Firebrowse web API for R. There wasn't any record in mRNA expression for 4 barcodes and I have excluded the. Resulting mRNA expression dataset contains 13332 records for three (NT, TP and TM) sample types.



I have 1097 records in clinical dataset and 13332 records in mRNA expression dataset (multiple lines for each barcode, one for each of the gene expression and z.score). Screen shot of clinical dta and mRNA data are as below.

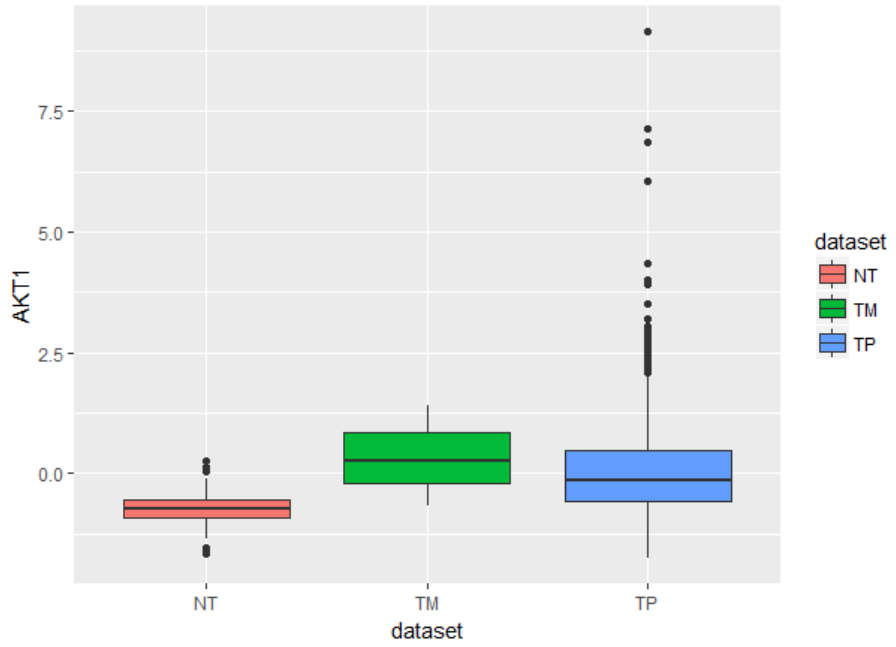
Clinical_data

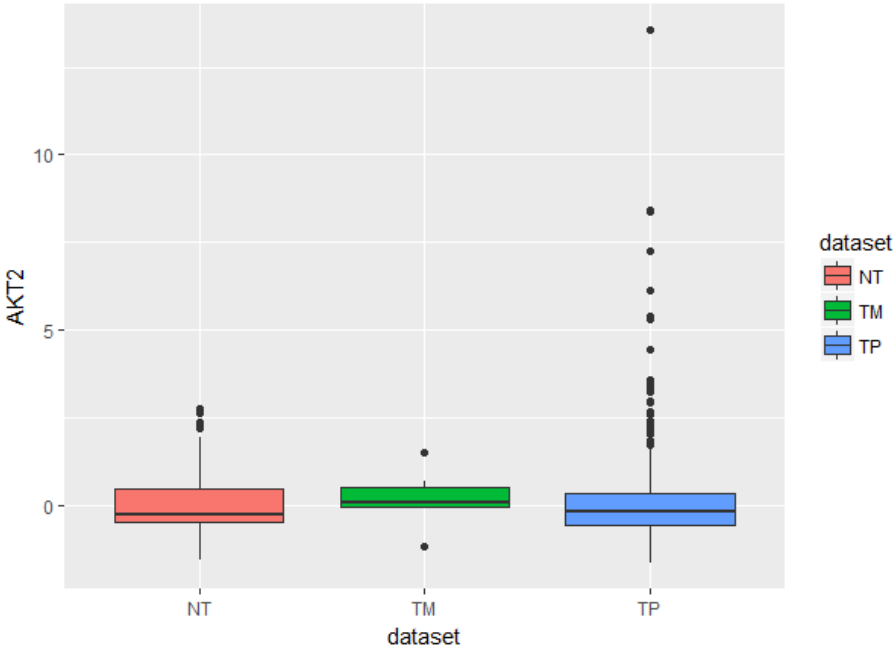
	A	B	D	E	F	J	K	L	M	N
1	tcga_participant_barcode	add	age_at_ini	anatomic_	anatomic_	axillary_ly	axillary_ly	bcr	bcr_canon	bcr_canon
2	TCGA-E9-A2JT		63	left upper outer quad		axillary lymph node di		nationwide children's hospital		
3	TCGA-BH-A0W4		46	left upper outer quadrant				nationwide children's hospital		
4	TCGA-BH-A0B5		40	left				nationwid	case matched express	
5	TCGA-AC-A3TN		75	right		sentinel lymph node b		nationwide children's hospital		
6	TCGA-BH-A0B3		53	right upper	right upper	sentinel lymph node b		nationwid	case matc	case matc
7	TCGA-A7-A0CD		66	left		sentinel node biopsy a		nationwide children's hospital		
8	TCGA-AN-A0G0		56	right upper inner quac		no axillary staging		nationwide children's hospital		
9	TCGA-E2-A10E		64	left		sentinel lymph node b		nationwide children's hospital		
10	TCGA-OL-A66K		72	right upper outer qua		sentinel node biopsy a		nationwide children's hospital		
11	TCGA-E2-A14N		37	right upper inner quac		sentinel lymph node b		nationwide children's hospital		
12	TCGA-5T-A9QA		52	left		no axillary staging		nationwide children's hospital		
13	TCGA-E2-A14V		53	left lower inner quadr		sentinel lymph node b		nationwide children's hospital		
14	TCGA-AC-A23H		90	right upper	right lower	inner quadrant		nationwid	case matc	case matc
15	TCGA-BH-A18J		56	right				nationwid	case matc	case matc
16	TCGA-A8-A06Z		84	right upper outer quadrant				nationwide children's hospital		
17	TCGA-A8-A06T		75	left upper outer quadrant				nationwide children's hospital		
18	TCGA-S3-AA10		65	right	right upper	axillary lymph node di		nationwide children's hospital		
19	TCGA-E9-A1N3		70	right upper inner quac		axillary lymph node di		nationwide children's hospital		
20	TCGA-LL-A6FP		90	right upper outer qua		no axillary staging		nationwide children's hospital		
21	TCGA-LL-A6FQ		77	right upper outer qua		axillary lymph node di		nationwide children's hospital		
22	TCGA-E9-A3HO		49	left upper outer quad		axillary lymph node di		nationwide children's hospital		
23	TCGA-A2-A0EY		62	right upper outer qua		sentinel lymph node b		nationwide children's hospital		
24	TCGA-A2-A0ET		58	right upper outer qua		axillary lymph node di		nationwide children's hospital		
25	TCGA-A2-A0ES		52	left upper outer quad		sentinel node biopsy a		nationwide children's hospital		
26	TCGA-A2-A0EN		70	right		sentinel node biopsy a		nationwide children's hospital		
27	TCGA-A2-A0EM		73	right upper inner quac		sentinel node biopsy a		nationwide children's hospital		
28	TCGA-A2-A3KD		47	right upper outer qua		sentinel lymph node b		nationwide children's hospital		

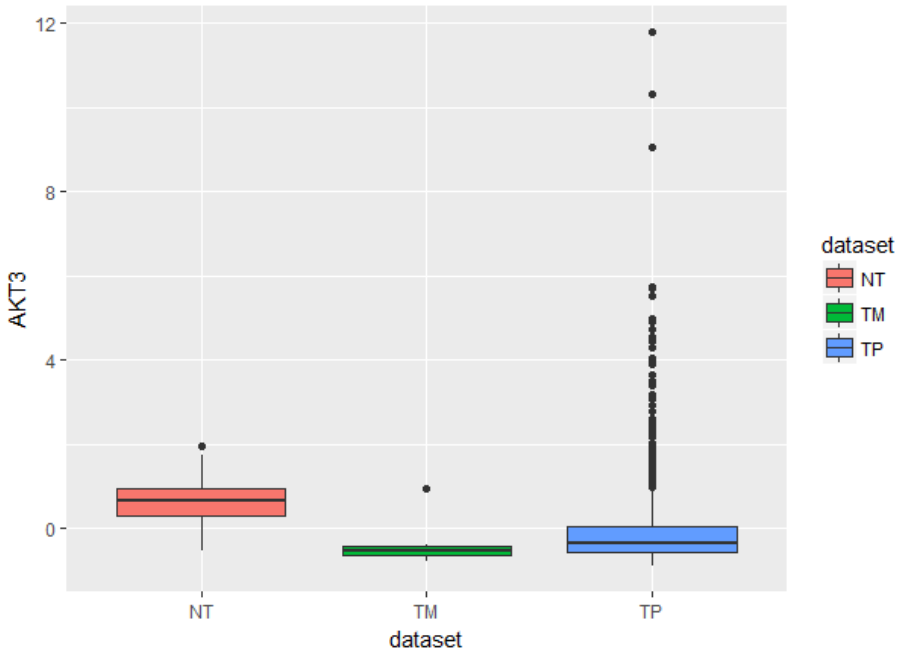
mRNA_Data

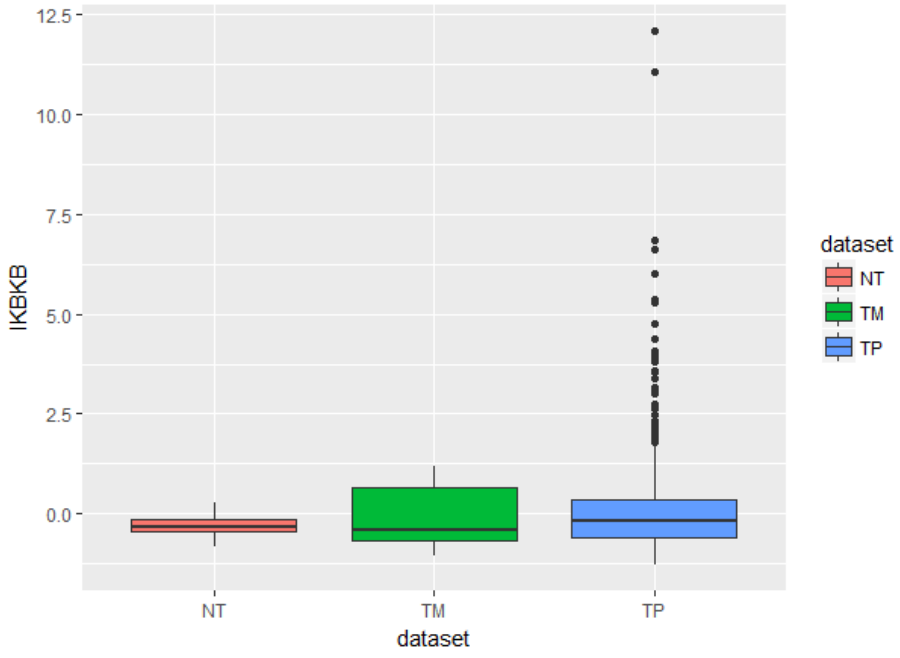
	A	B	C	D	E	F	G	H
1	tcga_participant_barcode	gene	express	z.score	cohort	sample_	protocc	geneID
2	TCGA-E9-A2JT	AKT3	9.703159	0.201575	BRCA	TP	RSEM	10000
3	TCGA-E9-A2JT	AKT2	11.52463	0.117994	BRCA	TP	RSEM	208
4	TCGA-E9-A2JT	AKT1	12.02042	-0.17704	BRCA	TP	RSEM	207
5	TCGA-E9-A2JT	IKBKB	11.38146	0.79141	BRCA	TP	RSEM	3551
6	TCGA-E9-A2JT	MYCN	6.610784	0.270942	BRCA	TP	RSEM	4613
7	TCGA-E9-A2JT	MYCL1	7.378777	-0.20641	BRCA	TP	RSEM	4610
8	TCGA-E9-A2JT	MYC	9.706831	-0.73509	BRCA	TP	RSEM	4609
9	TCGA-E9-A2JT	PRKAA2	7.787662	-0.24095	BRCA	TP	RSEM	5563
10	TCGA-E9-A2JT	RB1	10.61078	0.735284	BRCA	TP	RSEM	5925
11	TCGA-E9-A2JT	STK11	10.09191	0.193204	BRCA	TP	RSEM	6794
12	TCGA-E9-A2JT	TP53	10.80889	0.125726	BRCA	TP	RSEM	7157
13	TCGA-BH-A0W4	AKT3	9.022542	-0.20983	BRCA	TP	RSEM	10000
14	TCGA-BH-A0W4	AKT2	11.5155	0.103497	BRCA	TP	RSEM	208
15	TCGA-BH-A0W4	AKT1	12.10487	-0.05576	BRCA	TP	RSEM	207
16	TCGA-BH-A0W4	IKBKB	10.85356	0.113255	BRCA	TP	RSEM	3551
17	TCGA-BH-A0W4	MYCN	1.816641	-0.37301	BRCA	TP	RSEM	4613
18	TCGA-BH-A0W4	MYCL1	8.156488	0.195061	BRCA	TP	RSEM	4610
19	TCGA-BH-A0W4	MYC	11.38873	0.134225	BRCA	TP	RSEM	4609
20	TCGA-BH-A0W4	PRKAA2	8.19641	-0.11859	BRCA	TP	RSEM	5563
21	TCGA-BH-A0W4	RB1	10.49292	0.504125	BRCA	TP	RSEM	5925
22	TCGA-BH-A0W4	STK11	10.22139	0.410818	BRCA	TP	RSEM	6794
23	TCGA-BH-A0W4	TP53	10.55319	-0.20388	BRCA	TP	RSEM	7157
24	TCGA-BH-A0B5	AKT3	9.113836	-0.16525	BRCA	TP	RSEM	10000

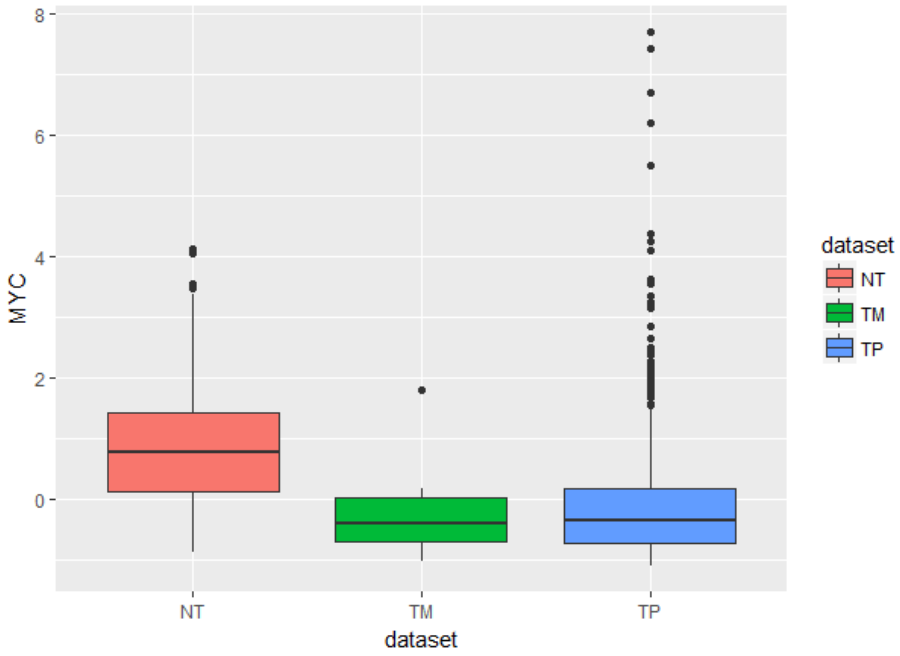
I have rearranged mRNA expression dataset and merged z_score of each of the gene to respective barcode of clinical dataset. I have plotted zscore distribution for each of the gene across three sample types using box plot and they are as below.

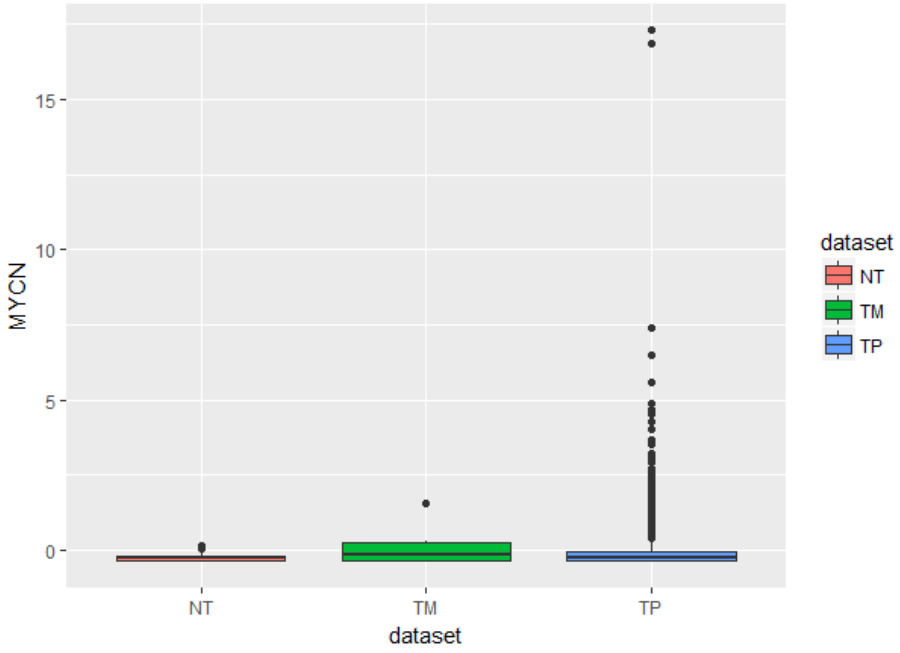


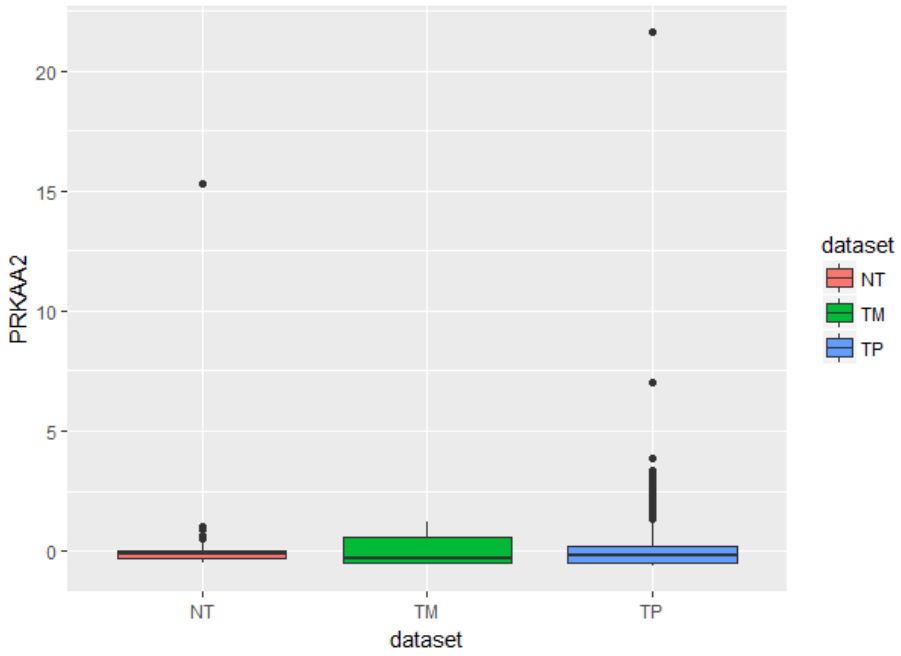


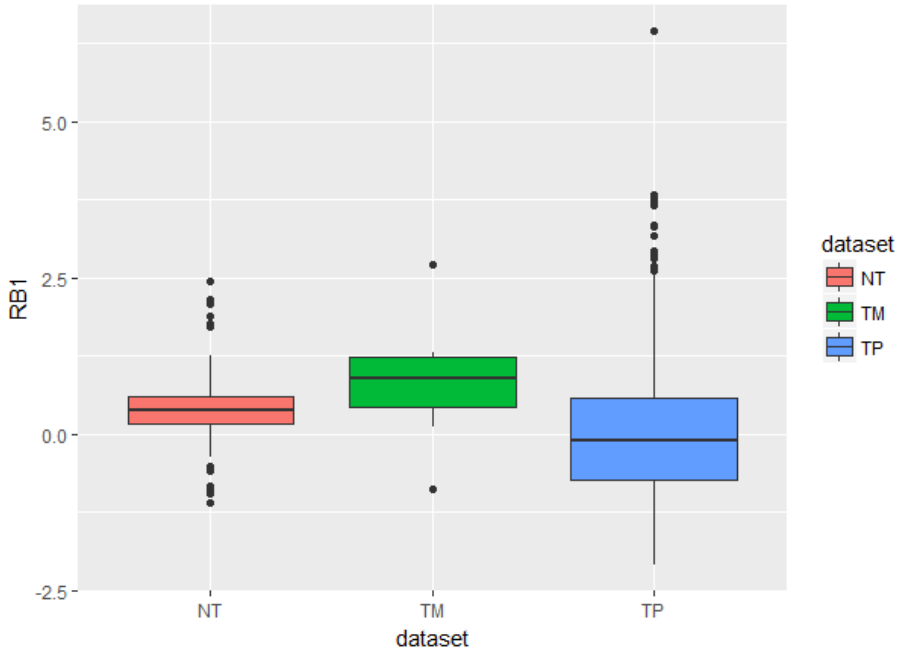

12
8
4
0
AKT3
NT
TM
TP
dataset
dataset
NT
TM
TP

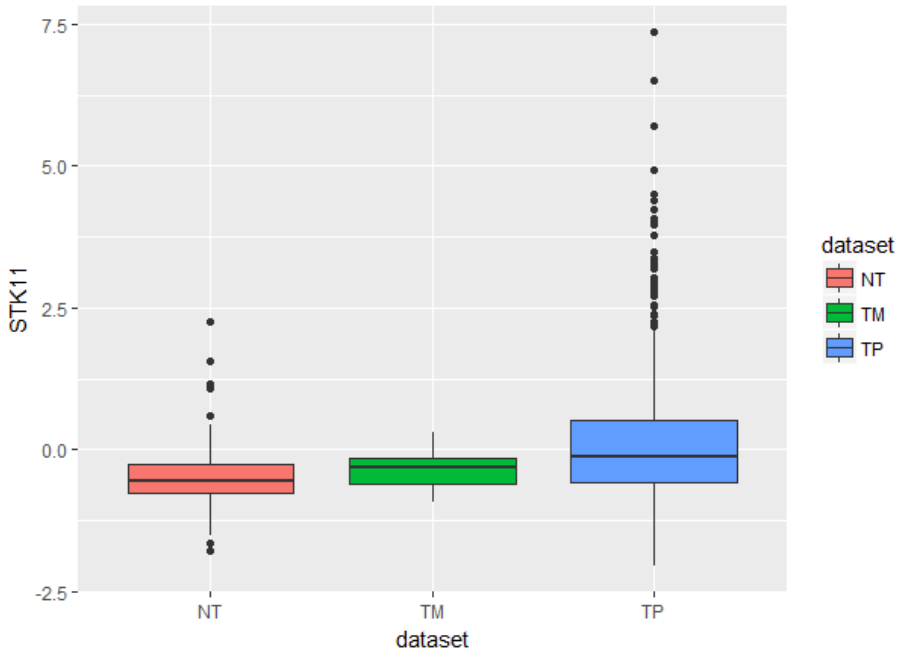

12.5
10.0
7.5
5.0
2.5
0.0
IKBKB
NT
TM
TP
dataset
dataset
NT
TM
TP

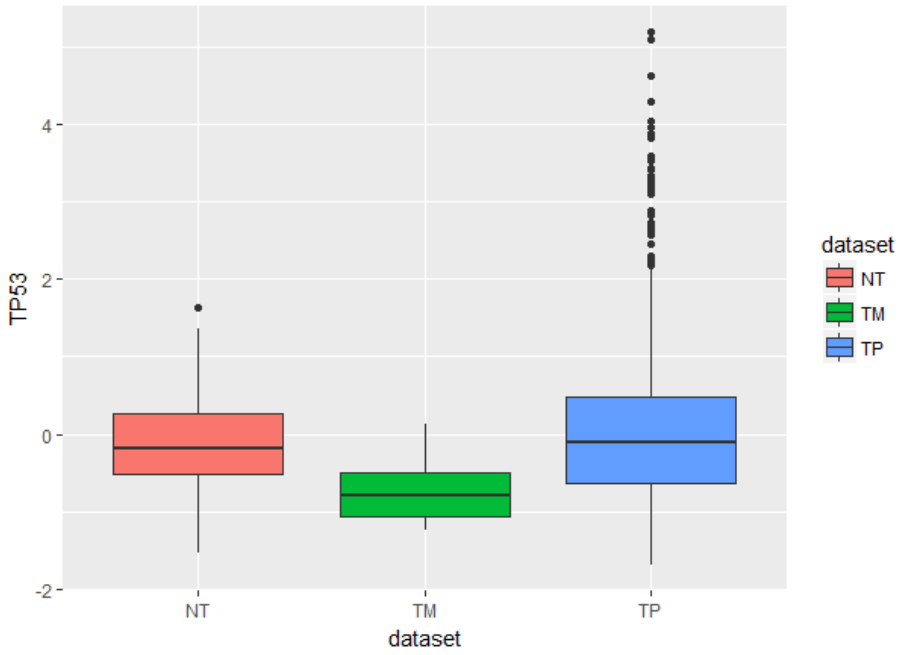

8
6
4
2
0
MYC
NT
TM
TP
dataset
dataset
NT
TM
TP


15
10
5
0
MYCN
NT
TM
TP
dataset
dataset
NT
TM
TP


PRKAA2
20
15
10
5
0
NT
TM
TP
dataset
dataset
NT
TM
TP


RB1
5.0
2.5
0.0
-2.5
NT
TM
TP
dataset
dataset
NT
TM
TP


7.5
5.0
2.5
0.0
-2.5
STK11
NT
TM
TP
dataset
dataset
NT
TM
TP


4
2
0
-2
TP53
NT
TM
TP
dataset
dataset
NT
TM
TP

Correlation plot across gene using zscore is as below. There is correlation between TP53, MYCL1, AKT3 and MYC as per below graph.

REFERENCES

[1] Giovanni, C. et al. Emerging landscape of oncogenic signatures across human cancers.
[2] Xi, Z. et al. Combining Gene Signatures Improves Prediction of Breast Cancer Survival
[3] The Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumors
[4] Cyriac, K. et al. Mutational landscape and significance across 12 major cancer types
[5] https://cancergenome.nih.gov/
[6] http://firebrowse.org/
[7] http://www.genecards.org/