GENT : Gene Expression across Normal and Tumor tissue

Overview

1. Introduction

Recent examples of successful cancer therapeutics such as Gleevec, Herceptin, and Iressa suggest that the concept of 'molecular targeted therapy' is applicable to human cancers of diverse tissue and genetic origin (Stuart and Sellers, 2009). 'Oncogene addiction' is a term to describe a phenomenon in which the growth and survival of tumors are impaired by the inactivation of a single oncogene (Weinstein and Joe, 2008). There are several established relationships between genetic alterations and corresponding targeted therapies, and efforts to identify further genetic alterations are underway. Mechanisms of genetic alterations include mutations (EGFR in lung cancer), translocations (BCR-ABL in chronic myeloid leukemia), and gene amplifications (ERBB2 in breast cancer) (Stuart and Sellers, 2009).

Interestingly, some addicted oncogenes are altered in only a subset of cancer patients. For examples, *ERBB2* is amplified and over-expressed in about 25-30% of breast cancer patients, whereas *EGFR* is mutated in about 20% of lung cancer patients. Cancer Outlier Profile Analysis (COPA) is a computational method that identifies gene expression profiles that are pathogenically over-expressed in only a subset of patients. *AGTR1* is an example of a potential target genes identified by applying the COPA method to the Oncomine database (Rhodes, et al., 2009).

A database with a large sample size is a great advantage when searching for genes over-expressed in only a subset of patients. For example, identifying genes over-expressed in 50 out of 1000 patients is easier and more reliable than identifying genes over-expressed in 2 out of 40 patients. Although the sample size of most individual gene expression studies rarely exceeds one thousand, a data set of nearly ten thousand samples (i.e., GeneSapiens database) can be created by a combined analysis of multiple data sets (Kilpinen, et al., 2008). Recent work has shown that analysis of a large microarray data set compiled from many data sets can reveal novel findings that are difficult to observe in the individual studies (Lukk, et al., 2010). For a combined analysis, data sets created by the Affymetrix platforms (i.e., U133A and U133plus2) offer several advantages. First, a majority of gene expression data sets have been created using the Affymetrix platforms. Second, many data sets are accompanied by raw CEL files so that users can pre-process them as they wish. We have collected human tissue gene expression data sets produced using the Affymetrix U133A and U133Plus2 platforms from public resources, and constructed a large-scale gene expression database of more than 40,000 samples.

2. DB description

More than 24,300 (U133plus2; 306 data sets) and 16,400 (U133A, 241 data sets) samples were collected and analyzed, and new data sets released on the Gene Expression Omnibus

(http://www.ncbi.nlm.nih.gov/geo/) and ArrayExpress (http://www.ebi.ac.uk/microarray-as/ae/) are monitored and updated every month (Barrett et al, 2009; Parkinson et al., 2009). The Expo (Expression Project for Oncology, http://www.intgen.org/expo/, also available as GSE2109) data set is the largest data set of cancer samples across diverse tissues.

Data sets according to platforms (U133plus2 and U133A) (From DB)

	U133Plus2	U133A	Total
Data sets	306	241	547
Samples	24,300	16,400	40,700
Probes	54,613	22,215	76,828
Total	79,219	38,856	118,075

3. Method

- Data preprocessing and normalization

Whenever CEL files were available (288/306 for U133plus2 and 191/242 for U133A), we pre-processed them using the MAS5 algorithm using the affy package (Gautier, et al., 2004). We chose the MAS5 algorithm because it is a single-array algorithm in which expression values are independent of other data. We then normalized each sample to a target density of 500. For data sets without CEL files but pre-processed by the MAS5 algorithm (18/306 for U133plus2 and 51/242 for U133A), we used expression measures downloaded from the web source and normalized them to a target density of 500.

- Data integration

We then classified each sample according to tissue and disease types. Most samples (~80%) were classified into either cancer or normal, but about 20% of samples were classified into other diseases including neurodegenerative diseases, immune-related diseases, and organ-specific diseases. We also collected expression data for more than 2500 samples comprising nearly 1000 different cancer cell lines across tissues, and processed them using the same method. The system implementation is based on an Apache web server, PHP scripts for data processing, R scripts for image production, and MySQL as a backend database.

4. Outlier detection

COPA (Cancer Outlier Profile Analysis) is a computational method that identifies gene expression profiles that are over-expressed in only a subset of patients (Rhodes et al. 2009). The original COPA method is based on comparison between 75th or 90th percentile to a normal median. We haven't implemented the original COPA method but instead use visualization by a box-plot so that users detect outliers across diverse tissues immediately.

5. Analysis of Laboratory Effects

Our database is composed of publicly available data sets in which laboratory effects are known to be strong. We assessed the impact of these effects following Lukk et al. (Lukk et al., 2010)'s analyses. We selected biological groups (with ten replicates or more) which contain at least two different laboratories. For U133A data sets, we selected 5089 samples of 92 biological groups produced in 93 laboratories. For each of the biological groups, we computed the average correlation coefficient between the assays from different laboratories within the same group. We also computed the average correlation coefficient between assays from the same laboratory but belonging to different biological groups. The comparison of the two similarity distributions showed that, as in Lukk et al. reported (Lukk et al., 2010), the biological effects were stronger than the laboratory effects (Figure XX). We got similar results with the U133Plus2 data sets, too.

6. Disclaimers

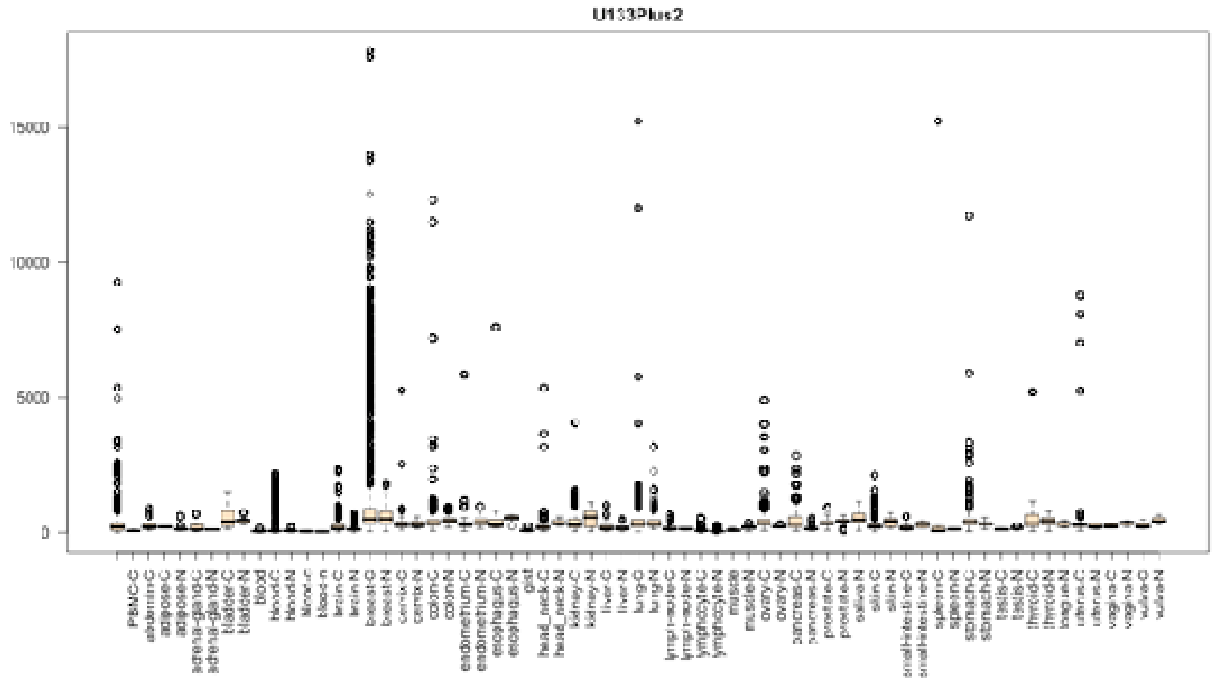
Our database is constructed from data sets collected from public resources such as Gene Expression Omnibus and Array Express. As these data were originated from many different laboratories in which experimental procedures may not be the same, we have NO guaranty that all the data have been generated by the same protocol and have the same high quality.

Anaysis

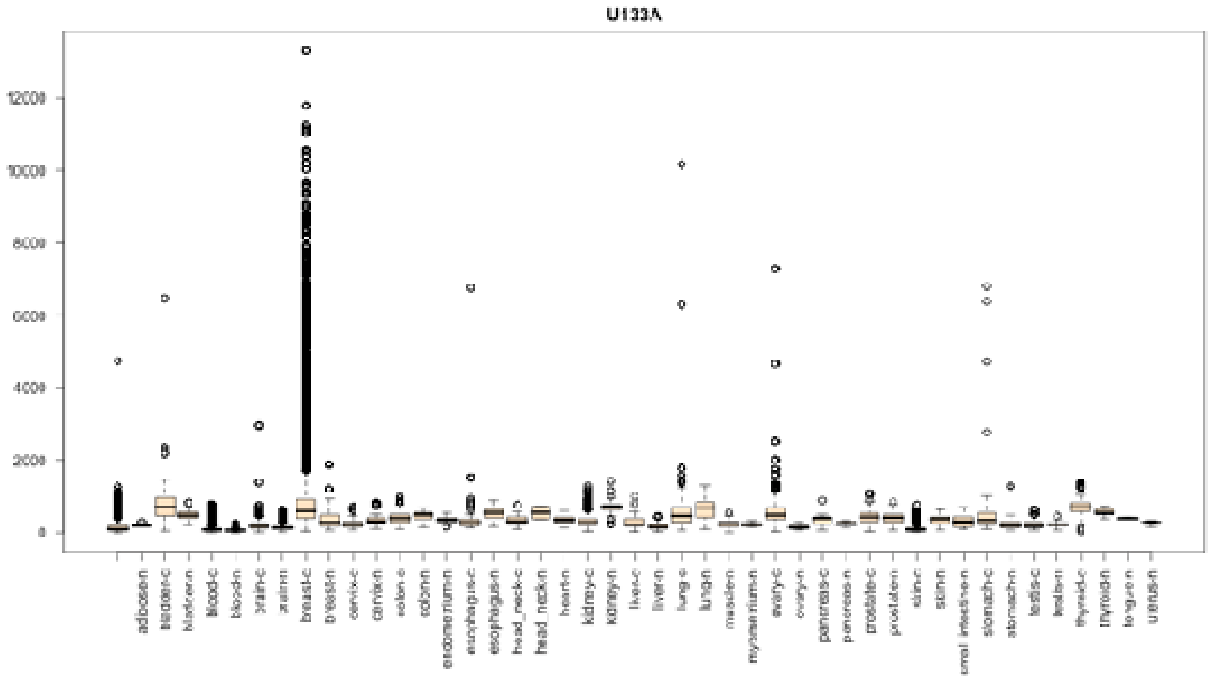
1. member of the ErbB protein family (*ERBB2*)

ERBB2 (HER2) is a member of the ErbB protein family, and is over-expressed in approximately 25-30% of breast cancer patients. Herceptin (trastuzumab) is a monoclonal antibody targeting ERBB2 over-expressing breast cancer patients. Over-expression of ERBB2 also occurs in other cancers such as lung, ovarian, and stomach cancer.

1) The pattern of ERBB2 in diverse tumor and normal tissues (U133Plus2 data set)

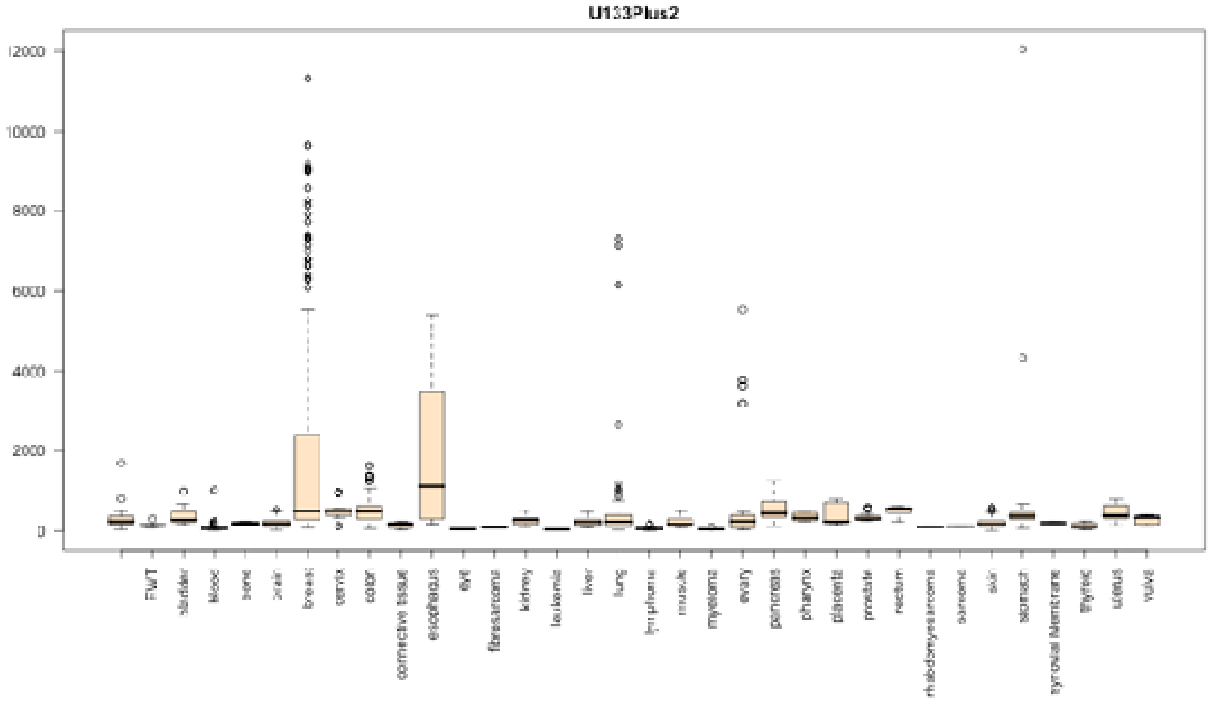


2) The pattern of ERBB2 in diverse tumor and normal tissues (U133A data set)

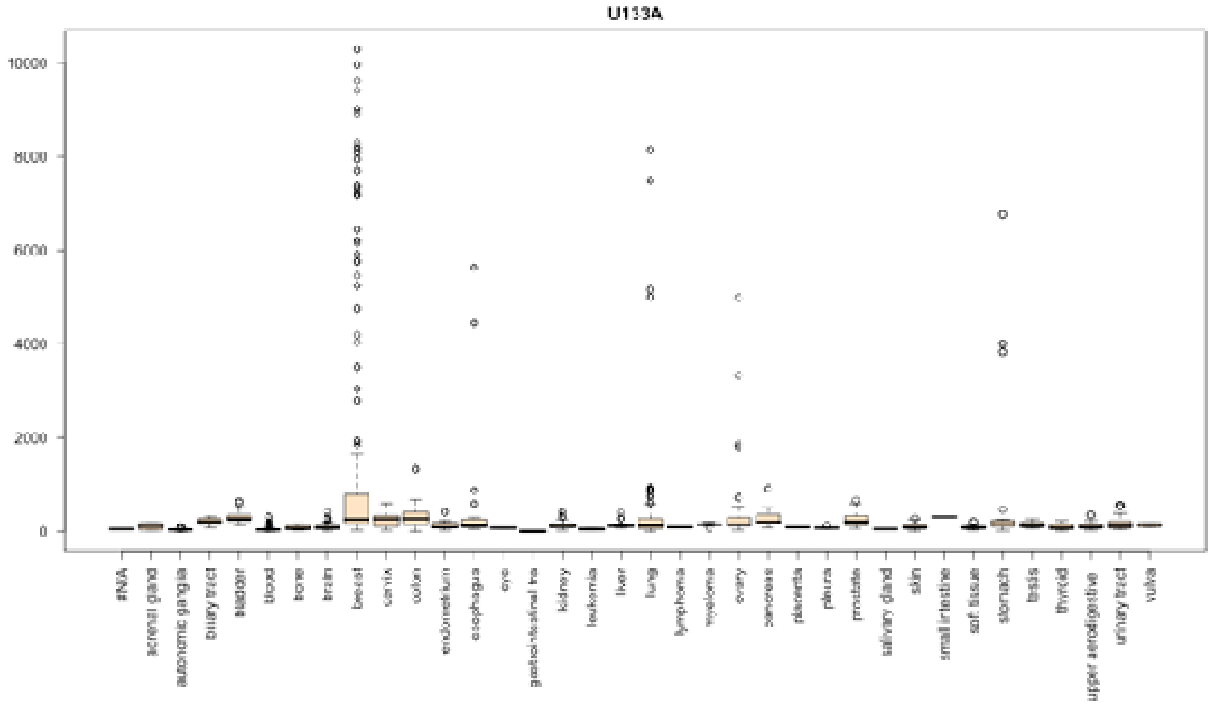


Both 1 and 2 clearly demostrate the over-expression of ERBB2 in a subsest of breast cancer patients as well as other cancer types such as lung, ovarian, and stomach cancers.

3) The pattern of ERBB2 in diverse cancer cell lines (U133Plus2 data set)



4) The pattern of ERBB2 in diverse cancer cell lines (U133A data set)

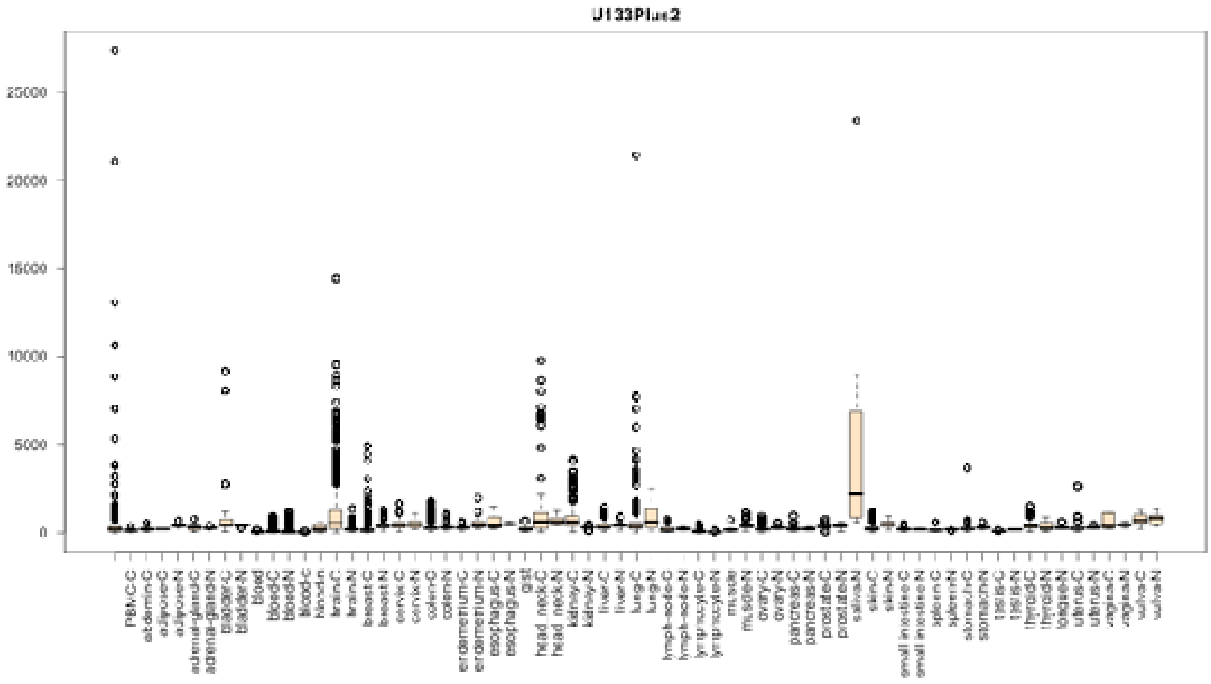


Interestingly, the pattern of ERBB2 expression in diverse tissues are recapitulated in cell lines, so a subset of breast, esophageal, lung, ovarian, and stomach cancer cell lines over-express ERBB2 (3 and 4). This information is especially useful when one selects cancer cell lines for in vitro siRNA (or shRNA) knockdown experiment.

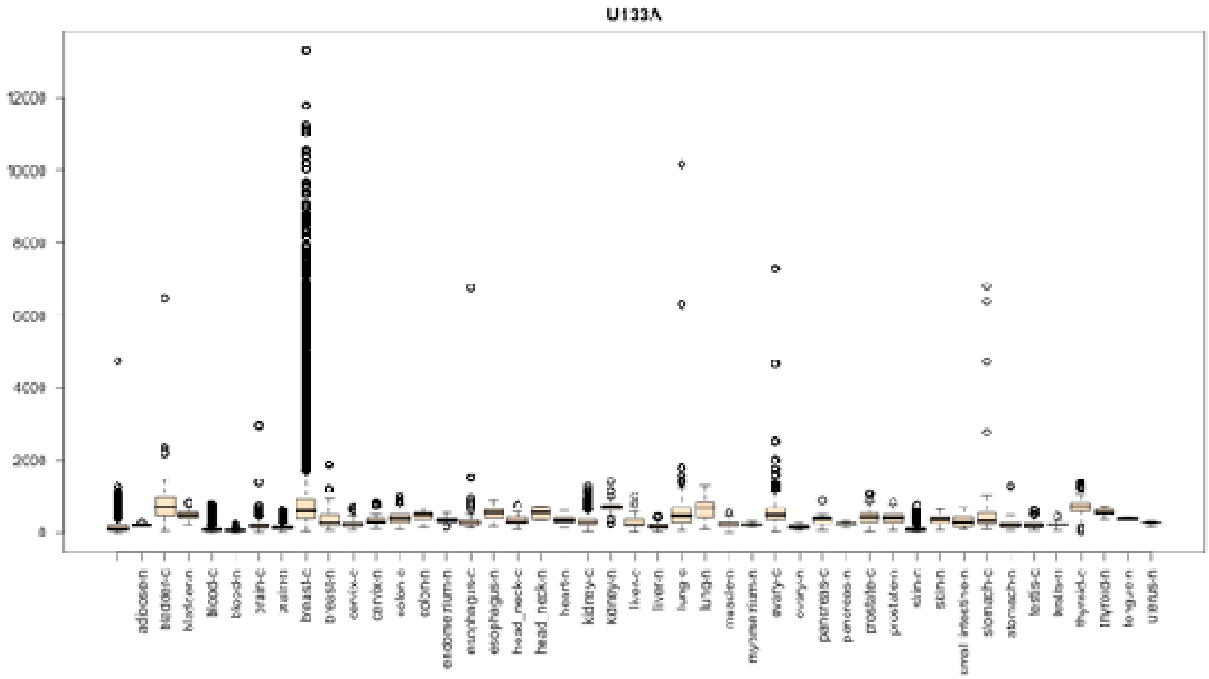
2. epidermal growth factor receptor (*EGFR*)

The epidermal growth factor receptor (EGFR) is a member of the ErbB family of receptors: EGFR, ERBB2, ERBB3, and ERBB4. Mutations and over-expression of EGFR are frequently observed in lung, colorectal, and brain tumors. Mutations, amplifications or misregulations of EGFR or family members are implicated in about 30% of all epithelial cancers. Iressa (gefitinib), erlotinib, and lapatinib are examples of small molecular kinase inhibitors targeting EGFR and cetuximab (Erbitux) and panitumumab are examples of monoclonal antibody inhibitors.

1) The pattern of EGFR in diverse tumor and normal tissues (U133Plus2 data set)

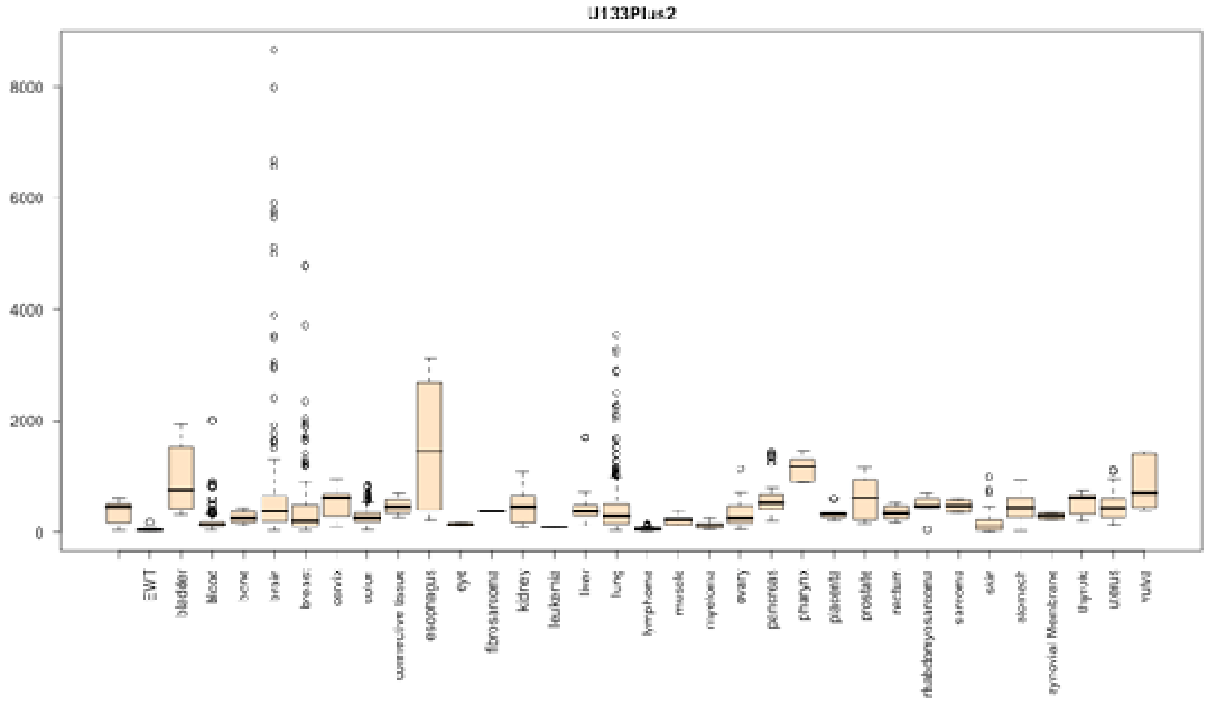


2) The pattern of EGFR in diverse tumor and normal tissues (U133A data set)

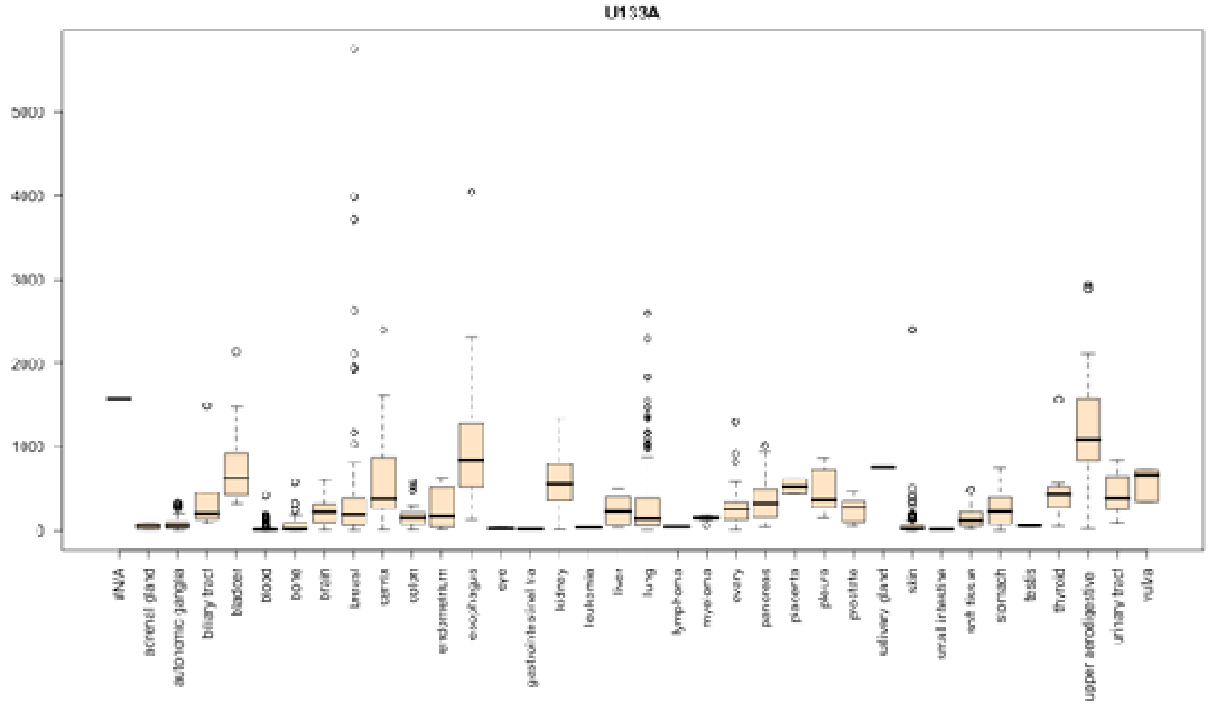


Again, the over-expression of EGFR in a subset of glioma and lung cancer patients are clearly demonstrated in both figures (U133Plus2 and U133A).

3) The pattern of EGFR in diverse cancer cell lines (U133Plus2 data set)



4) The pattern of EGFR in diverse cancer cell lines (U133A data set)



And the EGFR over-expression is recapitulated in a subset of glioma, breast, and lung cancer cell lines (3 and 4).

Information

1. Cancer Gene

The number of tissue samples according to tissue types (U133plus2 and U133A)

Tissue	U133Plus2		U133A		Total
	Cancer	Normal	Cancer	Normal	
abdomin	13	0	0	0	13
adipose	1	59	0	12	72
adrenal-gland	14	5	0	0	19
bladder	39	14	87	15	155
blood	2627	423	3180	1099	7329
brain	717	545	592	1627	3481
breast	1513	241	2635	91	4480
cervix	74	12	64	34	184
colon	1098	205	256	27	1586
endometrium	72	61	0	9	142
esophagus	13	9	24	28	74
GIST	0	0	0	0	64
head_neck	202	14	21	2	239
kidney	529	105	366	66	1066
liver	182	25	156	52	415
lung	599	160	684	454	1897
lymph-node	25	4	0	0	29
lymphocyte	414	144	0	0	558

Tissue	U133Plus2		U133A		Total
	Cancer	Normal	Cancer	Normal	
muscle	0	217	0	331	548
ovary	679	7	341	9	1036
pancreas	132	55	13	8	208
prostate	308	45	244	83	680
skin	238	28	499	59	824
small-intestine	13	6	0	0	19
stomach	250	42	46	18	356
testis	4	6	184	13	207
thyroid	62	25	44	25	156
tongue	0	11	0	4	15
uterus	155	12	0	24	191
vagina	3	5	0	0	8
vulva	21	14	0	0	35
heart	0	0	0	57	57
myometrium	0	0	0	24	24
small intestine	0	0	0	22	22
Total	**9997**	**2499**	**9436**	**4193**	**26189**

2. Cancer Cell Line

MWe also collected and processed cancer cell line data sets across diverse tissue types. Broad/Sanger Cancer Cell Line project (http://www.broadinstitute.org/cgi-bin/cancer/publications/pub_paper.cgi?mode=view&paper_id=189) and GSK's cell line project (https://array.nci.nih.gov/caarray/project/woost-00041/) provided the most abundant expression data sets.

Number of cell lines samples for each tissue type (U133A and U133P2) (From DB)

Tissue	U133plus2	U133A	Total
adrenal gland	0	2	2
autonomic ganglia	0	41	41
biliary tract	0	6	6

Tissue	U133plus2	U133A	Total
bladder	60	40	100
blood	449	142	591
bone	24	32	56
brain	200	76	276
breast	296	199	495
cervix	44	23	67
colon	203	56	259
endometrium	0	11	11
esophagus	24	25	49
eye	8	2	10
gastrointestinal tra	0	1	1
kidney	56	40	96
leukemia	0	4	4
liver	59	16	75
lung	587	325	912
lymphoma	76	1	77
myeloma	7	24	31
na	0	1	1
ovary	39	43	82
pancreas	77	18	95
placenta	18	2	20
pleura	0	6	6
prostate	24	18	42
salivary gland	0	1	1
skin	166	73	239
small intestine	0	1	1
soft tissue	0	19	19
stomach	71	24	95
testis	0	4	4
thyroid	24	13	37
upper aerodigestive	0	24	24

Tissue	U133plus2	U133A	Total
urinary tract	0	20	20
vulva	16	3	19
connective tissue	18	0	18
ewt	7	0	7
fibrosarcoma	1	0	1
muscle	19	0	19
pharynx	12	0	12
rectum	13	0	13
rhabdomyosarcoma	4	0	4
sarcoma	12	0	12
synovial membrane	6	0	6
uterus	44	0	44
Total	**2664**	**1336**	**4000**

Reference

1	Gautier, L., et al. (2004) affy--analysis of Affymetrix GeneChip data at the probe level, *Bioinformatics*, 20, 307-315.
2	Kilpinen, S., et al. (2008) Systematic bioinformatic analysis of expression levels of 17,330 human genes across 9,783 samples from 175 types of healthy and pathological tissues, *Genome Biol*, 9, R139.
3	Petrelli, N.J., et al. (2009) Clinical Cancer Advances 2009: major research advances in cancer treatment, prevention, and screening--a report from the American Society of Clinical Oncology, *J Clin Oncol*, 27, 6052-6069.
4	Rhodes, D.R., et al. (2009) AGTR1 overexpression defines a subset of breast cancer and confers sensitivity to losartan, an AGTR1 antagonist, *Proc Natl Acad Sci U S A*, 106, 10284-10289.
5	Stuart, D. and Sellers, W.R. (2009) Linking somatic genetic alterations in cancer to therapeutics, *Curr Opin Cell Biol*, 21, 304-310.

6	Weinstein, I.B. and Joe, A. (2008) Oncogene addiction, *Cancer Res*, 68, 3077-3080; discussion 3080.