Published in final edited form as:
J Proteome Res. 2013 January 4; 12(1): 45–57. doi:10.1021/pr300985j.

A Chromosome-Centric Human Proteome Project (C-HPP) to Characterize the Sets of Proteins Encoded in Chromosome 17

Suli Liu[1], **Hoguen Im**[2], **Amos Bairoch**[3], **Massimo Cristofanilli**[4], **Rui Chen**[2], **Eric W. Deutsch**[5], **Stephen Dalton**[13], **David Fenyo**[6], **Susan Fanayan**[7], **Chris Gates**[10,14], **Pascale Gaudet**[3], **Marina Hincapie**[1], **Samir Hanash**[8], **Hoguen Kim**[9], **Seul-Ki Jeong**[10], **Emma Lundberg**[15], **George Mias**[2], **Rajasree Menon**[11], **Zhaomei Mu**[4], **Edouard Nice**[12], **Young-Ki Paik**[10,13], **Mathias Uhlen**[15], **Lance Wells**[16], **Shiaw-Lin Wu**[1], **Fangfei Yan**[1], **Fan Zhang**[1], **Yue Zhang**[1], **Michael Snyder**[2], **Gilbert S. Omenn**[11], **Ronald C. Beavis**[15], and **William S. Hancock**[1,7,13,*]

[1]Barnett Institute and Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115, USA [2]Stanford University, Palo Alto, California, USA [3]Swiss Institute of Bioinformatics (SIB) and University of Geneva, Geneva, Switzerland [4]Fox Chase Cancer Center, Philadelphia, PA, USA [5]Institute for System Biology, Seattle, Washington, USA [6]School of Medicine, New York University, USA [7]Department of Chemistry and Biomolecular Sciences, Macquarie University, Sydney, NSW, Australia [8]MD Anderson Cancer Center, Houston, TX, USA [9]Yonsei University College of Medicine, Yonsei University, Seoul, Korea [10]Yonsei Proteome Research Center, Yonsei University, Seoul, Korea [11]Departments of Computational Medicine & Bioinformatics, Internal Medicine, Human Genetics, and School of Public Health, University of Michigan, Ann Arbor, MI, USA, and Institute for Systems Biology, Seattle, WA, USA [12]Monash Antibody Technologies Facility, Monash University, Clayton, VIC 3800, Australia [13]Department of Integrated Omics for Biomedical Science, Yonsei University, Seoul, Korea [14]Compendia Biosciences Inc, Ann Arbor, MI, USA [15]Science for Life Laboratory and Albanova University Center, Royal Institute of Technology (KTH), Stockholm, Sweden [16]Complex Carbohydrate Research Center, University of Georgia, 315 Riverbend Rd. Athens, Georgia

Abstract

We report progress assembling the parts list for chromosome 17 and illustrate the various processes that we have developed to integrate available data from diverse genomic and proteomic knowledge bases. As primary resources we have used GPMDB, neXtProt, PeptideAtlas, Human Protein Atlas (HPA), and GeneCards. All sites share the common resource of Ensembl for the genome modeling information. We have defined the chromosome 17 parts list with the following information: 1169 protein-coding genes, the numbers of proteins confidently identified by various experimental approaches as documented in GPMDB, neXtProt, PeptideAtlas, and HPA, examples of typical data sets obtained by RNASeq and proteomic studies of epithelial derived tumor cell lines (disease proteome) and a normal proteome (peripheral mononuclear cells), reported evidence

[*]Corresponding author: William S. Hancock, Barnett Institute and Department of Chemistry and Chemical Biology, Northeastern University, Boston, MA 02115, USA wi.hancock@neu.edu Telephone: 617-373-4881 Fax: 617-373-8795.

Supporting Information Available: This material is available free of charge via the Internet at http://pubs.acs.org.

of post-translational modifications, and examples of alternative splice variants (ASVs). We have constructed a list of the 59 'missing' proteins as well as 201 proteins that have inconclusive mass spectrometric (MS) identifications. In this report we have defined a process to establish a baseline for the incorporation of new evidence on protein identification and characterization as well as related information from transcriptome analyses. This initial list of 'missing' proteins that will guide the selection of appropriate samples for discovery studies as well as antibody reagents. Also we have illustrated the significant diversity of protein variants (including post-translational modifications, PTMs) using regions on chromosome 17 that contain important oncogenes. We emphasize the need for mandated deposition of proteomics data in public databases, the further development of improved PTM, ASV and single nucleotide variant (SNV) databases and the construction of websites that can integrate and regularly update such information. In addition, we describe the distribution of both clustered and scattered sets of protein families on the chromosome. Since chromosome 17 is rich in cancer associated genes we have focused the clustering of cancer associated genes in such genomic regions and have used the ERBB2 amplicon as an example of the value of a proteogenomic approach in which one integrates transcriptomic with proteomic information and captures evidence of co-expression through coordinated regulation.

Keywords

Chromosome-Centric Human Proteome Project; Chromosome 17 Parts List; ERBB2; Oncogene

Introduction

A new scientific initiative, the Chromosome-Centric Human Proteome Project (C-HPP) of the Human Proteome Organization, has a 10 year goal of characterizing the 'parts list' of the entire human proteome encoded by the approximately 20,300 human protein-coding genes.[1,2] We believe that integration of proteomics data into a genomic framework will promote a better understanding of the relationship of the transcriptome to the proteome and facilitate international collaborations with different national teams volunteering for an individual chromosome. In this manner a group of primarily US-based scientists have decided to study chromosome 17 and to characterize the full set of proteins coded by this chromosome as well as identify the major variants. The reason for selection of this chromosome was based on the presence of the driver oncogene, ERBB2 as well as the close association of a significant number of genes present on chromosome 17 with cancer. In addition, our team has developed a close association with the Australian and New Zealand scientists who are studying chromosome 7 which contains the oncogene EGFR, which together with ERBB2 forms a heterodimer complex which results in receptor kinase activation and oncogenic signaling. We will, therefore, report in this publication on the current status of the proteogenomic parts list of chromosome 17 and discuss future steps in our part of the C-HPP initiative[1].

The DNA sequence of chromosome 17 was most recently defined in 2006[3] and chromosome 17 contains 78,839,971 bases or 2.8% of the euchromatic genome. In RNA-sequencing studies it was noted that there is an average of 5 distinct transcripts per gene locus and approximately 75% with at least two transcripts, as well as some 274 pseudogenes[3]. Chromosome 17 was also found to have some unusual properties. It contains the second highest gene density of all chromosomes (16.2 genes per Mb) and is enriched in segmental duplications and non-allelic homologous recombinations (NAHR). Non-allelic homologous recombination can occur during meiosis in which crossing over between strands results in duplication or deletion of the intervening sequence[3]. Such deletion or duplication of regions of the genome may be related to the association of human chromosome 17 with a wide range of human diseases e.g. microdeletion disorders which occur in the 17p12 and 17p13 regions, loss of 17p in CNS tumors, a loss of heterozygosity (LOH) which involves the BRCA1 gene (17q21), 17q gain and isochromosome formation in neuroblastomas, and translocations between chromosome 17 and chromosome 5, 9,11,15 or 22 in various diseases[4].

Results and Discussion

Section 1 Background to chromosome 17 and its unique properties, especially related to cancer biology

Chromosome 17 has a strong association with cancers and is extensively rearranged in at least 30% of breast cancer tumors. Whereas the short arm undergoes frequent losses, the long arm has complex combinations of overlapping gains and losses[5]. Studies of the transcriptome map revealed regions (17 p11, p13, q11, q12, q21, q23 and q25) with higher expression levels in specific chromosomal regions in 10 tumor tissue types[6]. An increase in gene copy numbers for the region from 17q22 to 17q24 was observed in a large set of cancer cell lines and primary tumors by comparative genomic hybridization and cDNA arrays[7] and was related to amplification of ERBB2 and collinear genes. In fact, chromosome 17 contains many regions with cancer-associated genes (see Table 1), including such prominent genes as TP53 (DNA damage response/usually called tumor suppressor), BRCA1 (breast cancer), NF1 (neurofibromatosis) and ERBB2 (breast cancer). To construct a more complete list of cancer associated genes we have integrated information from several web sites (primarily Sanger, Waldman and GeneCards, see legend to Table 1) and have identified 44 such genes on chromosome 17. In this table we also explore the tendency of genes that are strongly associated with cancer to originate in transcriptionally active regions (high gene density) and to be clustered with other cancer related genes[4,8]. In chromosome 17 these associations are indeed observed; only 8 of the 44 oncogenes occur in a region with a gene density of less than 30 genes per Mb and all oncogenes had >5 other cancer associated genes in proximity. In addition, regions identified with high expression in tumor studies[6] contained significant numbers of cancer genes listed in this Table: p11 (2 including FLCN), p13 (9, TP53), q11 (NF1), q12 (4, ERBB2), q21 (13, BRCA1), q23 (4, DDX5) and q25 (9, GRB2).

Section 2 List of Protein Coding Genes as Baseline for C-HPP and corresponding transcriptomic and proteomic information

Our current knowledge of the proteogenome of chromosome 17 was aggregated by information stored in the following resources: neXtprot[9], Uniprot[10], Unipep[12], GPMDB[14], PeptideAtlas[11,16], Genecards[15], Oncomine[13] and Human Protein Atlas (see this issue). The data sets of GPMDB(release 2012/07/01) and PeptideAtlas(release 2012/09) are based on an aggregation of curated protein identifications by mass spectrometry that have been deposited in the public domain (PRIDE, PeptideAtlas, Tranche) and then undergone standardized reanalysis of the mass spectra. Direct deposition at neXtProt(Release 2012/09/11) is based on the annotation resources of SwissProt and UniProt[10], including literature curation. Thus the different major databases represent alternative snapshots of the information flow into proteomics ranging from experimental data sets to reviewed publications to epitope-based antibodies and immunohistochemistry at HPA (release 2012/09/12). The designation of the status of protein-coding genes is based on the Ensembl genome browser (current version 68). For chromosome 17 there are 1169 such genes with the following information in Uniprot; 861, 269 and 40 with the designation of protein or transcript evidence or with uncertain status (green, yellow, and red, respectively in Figs 1 and 3). For the HPP, our baseline accepts only those protein characterizations with the highest-grade identifications, not just those inferred from transcripts, and recognizes the propensity to false discovery and lack of confirmation of many reported findings. The number of such highest-grade identifications for this chromosome is as follows: 601 (51%, neXtProt gold), 824 (70%, GPMDB, green) and 745 (64%, PeptideAtlas, 1% FDR for proteins) of the number of protein coding genes, respectively. As examples of the state of knowledge at the level of proteomics, the numbers of protein identifications with lower status listed in GPMDB and PeptideAtlas are, respectively, medium (49 and 43), low probability (201 and 190) and ‘missing’ or black as 95 in GPMDB. In Protein Atlas there are antibodies corresponding to 725 genes on chromosome 17 of which 503 are of high quality (HPA score medium or high) and in addition the number of available polyclonal and monoclonal antibodies listed in antibodypedia and Labome (see figure 3 legend) are currently 900 and 416 respectively.

One goal of the C-HPP initiative is to promote the integrated analysis of multiple ‘omics platforms, starting with use of RNA-Seq studies to guide deeper proteomics. The HPP initiative has the overall goal of both defining the protein parts list and establishing the biology/disease context of such information; we wish to capture in this discussion the significant amount of ‘parts’ information that could be fed into biology with such an integrated approach. As a starting basis, with common ‘shotgun’ proteomics approaches one could expect only about 50 out of a total of 1169 predicted proteins to be identified on chromosome 17 in a proteomic study with 1000 identifications (4.5% of the total genome). The number of identifications of proteins coded by genes present in chromosome 17 was 140 in the disease study (two cell lines, triplicates, LTQ-FTMS) and 237 in normal sample (peripheral mononuclear cells[17], a very deep study with 164 serial analyses, orbitrap) vs. 601 and 730 transcripts (disease and normal respectively). In proteomic studies the number of observed proteins will be related to the approach used (sample fractionation steps, type of mass spectrometer) as well as the number of replicates and size of the sample set but in both cases the depth of the transcriptomic finding was much deeper than the proteomics and underlines the importance of promoting the adoption of improved analytical procedures in proteomics[18]. The disease implications of this comparison will be the subject of a separate manuscript in which we discuss the selection of appropriate control or normal samples for a disease biomarker study.

To further illustrate the data sets generated by a combined proteomic/ transcriptomics study and compare with information listed in Uniprot and GPMDB, Fig. 1 shows the region around the important and well-studied oncogene ERBB2. Again the RNASeq data sets are more extensive (11 and 12 transcripts for the 20 genes flanking ERBB2, respectively) vs. proteomic coverage (5 and 6 respectively). Uniprot has annotated 5 genes in this region only at the level of transcript (STAC2, NEUROD2, PGAP3, ZPBP2, and LRR3C) and one of these was not observed at the transcript level in our studies (LRR3C). In the aggregated data of GPM one protein from the gene STAC2 was observed with medium level probability (yellow) and 4 with low probability (red) while one had no evidence (LRR3C).

Section 3 The location of protein families on chromosome 17

As is shown in Fig. 2 there are several regions on the chromosome with clusters of genes in protein families. Genes in human families can exhibit close clustering on a single chromosome and presumably arose by tandem gene duplication[19], such as the growth hormone family (5 genes) at 17q23, CD300 (7 genes, 6 clustered at 17q25.1) and Schlafen family of 5 growth regulatory genes at 17q12. Alternatively, gene families may be of a complex type and dispersed and consist of multiple gene clusters at different chromosomal locations, such as the olfactory receptors at 17p13.2, 13.3 where 12 of a large family of 398 members are present. In this case the evolution process may have involved a mixture of gene and genome duplication events and the sequences may be divergent. Other examples of gene families substantially clustered on chromosome 17 include cytokines, chemokine ligands, keratins, keratin - associated proteins, homeobox and chromobox proteins. The number of protein coding genes in each band is listed in red in the figure and interestingly band 17q21.2 contains 109 genes with 50% comprised of either keratin or keratin-associated proteins (28 and 25, respectively). In the future we plan to explore the challenge of characterizing closely homologous proteins that may be very difficult to identify unambiguously by proteomics, given limited sequence coverage that is obtained in a proteomics experiment.

Section 4 What are the ‘missing or black proteins’?

To guide the search for ‘rare’ tissue and cell lines samples and experimental protocols we have developed a list of the 59 ‘missing’ proteins from chromosome 17. The list is limited to only those proteins that do not have any proteomic identifications in neXtProt, PeptideAtlas or GPM. The last column in Figure 3 shows the availability of antibody evidence for protein expression from HPA and only 4 genes have good antibody evidence (MYH4, HOXB1, CD300C, CD300LD). There is, in addition, significant number of proteins with only preliminary evidence in the databases. For example, the number of low probability identifications listed in GPMDB and PeptideAtlas for chromosome 17 are 201 and 190 respectively. A further example of incomplete information is provided by the important region of 20 genes around ERBB2, where 1 gene product has not been identified and 4 only with low levels of evidence (Figure 1). We will concentrate on the ‘missing’ proteins initially, but in the future we will perform targeted studies on suitable tissues that are rich sources for poorly characterized proteins and deposit additional datasets in the public databases to improve the confidence of these identifications. As shown in Fig. 3 tissues for targeted analyses can be identified through transcript expression data. The development of a list of 'missing' proteins will also allow the identification of situations where the lack of identification is due to a technical issue (poor enzymatic cleavage steps or unsuitable protein/peptide physical properties) and also identify cases where the gene model itself is problematic or the protein nomenclature is confounded due to synonyms.

The process we used to refine the list of ‘missing’ proteins was as follows:

1. Eliminate any faulty gene annotation of non-protein coding genes (2 were eliminated from the chromosome 17 list and both were also identified as red in the Uniprot evidence list).

2. Identify ‘missing’ proteins that had abundant RNASeq evidence (and in some cases proteomic data) from our experimental data sets (epithelial cancer cell lines, peripheral mononuclear cells). In some cases we identified nomenclature problems where a ‘missing’ proteins was in fact reported in the proteomic databases under alternate gene symbols with good quality identifications. While our cell line data shown in Fig. 3 did not show any highly credible protein IDs some did show significant levels of transcript (see Fig.3 yellow or green transcript levels for genes RNF43, STRADA, ARL16). Also one can use gene and transcript data from other sources (obtained from GeneCards in this figure legend) to identify target tissues for follow up RNASeq and proteomic studies. For example, we have initiated a study on nasal epithelial cells to identify missing olfactory receptors. In addition, brain and hair cortex look to be promising samples for additional studies. Brain has long been known to express a much higher proportion of single-copy DNA expressed in mRNA than in liver, kidney, and spleen[20].

3. Obtain suitable antibodies specific for the ‘missing’ proteins: in the case of chromosome 17 there are 5 ‘missing’ proteins with no antibody availability. Our next step is to prepare suitable monoclonal antibodies in collaboration with the antibody resources of the initiative. These antibodies will be used to confirm the proteomic identifications in Western blots as well as for affinity isolation steps to facilitate the identification of protein isoforms.

Section 5 Splice Variant analysis of Chromosome 17 genes based on RNA-Seq data from six ErbB2+ cancer cell lines

Alternative splicing and post-translational modifications in higher eukaryotes increase the diversity of protein products derived from a single genetic locus and enable regulation of cellular and developmental processes. Across various cancers, examples of aberrant splicing events include alternative splice sites, alternative promoters, exon skipping, retained introns, and inclusion of presumed 5’ or 3’ untranslated regions (UTRs). The translated protein products of the alternatively spliced transcripts of a gene may play different roles in cancer mechanisms as there are various studies showing distinct opposite functions for the variants from a single gene[21,22].

Activation of the ErbB2 receptor signaling pathways has been shown to increase cancer metastasis[23]. We studied alternatively spliced transcripts (ASTs) expressed as distinct RNA-Seq reads from transcript-specific exons in the following six ERBB2+ cancer cell lines: two colorectal (LIM2405, LIM1899), two gastric (KATOIII, SNU16) and two breast (SUM149, SUM190). Across these 6 cancer cell lines, we identified 195 distinct ASTs from 144 genes;

46 of the 144 genes had more than one alternative transcript expressed (Table 2). We compared the differential expression of the transcripts in these cell lines based on the FPKM (fragments per kilobase of exon per million fragments mapped). Interesting observations include the following:

1. Seven distinct ASTs of septin 9 (SEPT9) were identified, with strikingly different expression levels of these ASTs across the six cell lines (Figure 4(a)); sept9 epsilon is the isoform most expressed, by far. Altered expression of sept9 is observed in several carcinomas[24].
2. We identified reads from exons specific for the variant of ERBB2 which translated to the shorter protein variant (ENSP00000385185) (Figure 4(b)).
3. The longest AST variant of each of three genes (CDK12, FBXL20, GRB7) that are located quite near the ERBB2 gene on chromosome 17 was identified (Figure 4(c)).
4. Over-expression of the shorter variant of PPP1R1b was observed in the KATOIII colorectal cancer cell line (Figure 4(d)).

Section 6 Additional Protein variants resulting from Alternative Splice and Single Nucleotide Variants and Post-translational modifications (PTMs)

We illustrate the diversity of protein isoforms for three regions on chromosome 17 that are adjacent to three important oncogenes, ERBB2, TP53 and BRCA1 (see Figure 5 and in supplementary Figures 1 and 2, 10 genes on either side). In this figure we list the number of alternative splice variants (ASVs, red circle) and single nucleotide variants (SNVs, purple circle) resulting from the transcription of missense SNPs. Examples of genes with a significant number of variants in the region around the four important oncogenes include ERBB2 with 4 and 88, IKZF3 (15, 7), TP53 (9, 1394), and BRCA1 (6, 320) ASVs and SNVs, respectively. In this figure we also list the four major PTMs (neXtProt data) in boxes with separate listing for phosphorylated serine and threonine vs. tyrosine and for N- vs. O-glycosylation (see figure legend). At the level of the entire chromosome 17 with 1169 protein coding genes, the number of proteins with at least one known post-translational modification is 526 (approximately 45%), with the following major categories (426 phosphorylations; 185 acetylations; 51 glycosylations; 10 methylations; 8 palmitoylations; 2 myristoylations; many proteins with several identifications). In the cases of ERBB2, TP53 and BRCA1 regions, 9, 6, and 9 of the 20 surrounding genes have no information on PTMs, respectively, which is similar coverage to that of the entire chromosome. The 5 genes in the ERBB2 region with no PTM information in GPMDB are listed in Uniprot with only transcript evidence (STAC2, NEUROD2, PGAP3, ZPBP2, and LRR3C) while two other genes are observed with very low expression levels in many cancer studies (PNMT, TCAP, see below for discussion). These examples illustrate that there is a lack of PTM information even in well-studied regions of the genome, despite recent advances in glycomics/ proteomics experimental methods[25]. A further and significant analytical and bioinformatic challenge is the site identification and degree of occupancy for phosphorylated residues and the heterogeneity of N-glycosylated structures; this summary does not attempt to capture this level of data complexity.

An important future goal of proteomics, as well as functional genomics, is to identify which of the potential protein variants that are identified at the genome or transcriptome level are observed at the level of the proteome. An example of this process is the characterization of the numbers of ASVs of ERBB2 observed in the proteome that is listed as6 in Ensemble (Release 2012/07) and 4 in neXtPort. There is more complexity to be revealed, however: the total number of ASTs listed in Ensembl is 14 with the following amino acid lengths; 1255, 1240, 1225 (3 forms), 1055, 979, 603 (2), and <252 (5 forms). The shorter variants have not be observed at the protein level and only 4 ASVs are listed in the consensus CDS protein set (an NCBI collaborative effort to identify the well annotated core set of human protein coding regions); these variants are listed with proteomics information in neXtProt (amino acid lengths of 1255, 1225 (3)). A larger set of 6 protein forms is listed in GeneCards (1255, 1240, 1225 (2) and 979) and same set of 6 is listed in GPM with MS identifications.

One could also expect that the site and nature of PTMs could be altered in a protein variant. Since there is little information on the MS characterization of SNVs in the human proteome, we have used data from the NCBI dbSNP database to capture potential polymorphism sites through ENSEMBL BioMart[26]. Again using ERBB2 as an example, we examined the PTM information for 5 ASVs listed for ERBB2 in GPMDB. For ENSP00000269571, phosphorylation sites are *T701*, Y735, T862, Y877, S998, Y1023, S1051 (S/Y for SNV), S1054, S1073, S1083, S1107, S1151, T1166, S1174, S1214, T1240, Y1248; for ENSP00000385185, phosphorylation sites are *T671*, Y705, T832, Y847, S968, Y993, S1021 (S/Y for SNV), S1024, S1043, S1053, S1077, *Y1109*, T1136, S1144, T1210, Y1218; for ENSP00000443562, sites are *T671*, Y705, S968, Y993, S1024, S1048, S1121, S1144, T1210; for ENSP00000446466, sites are *T686*, Y720, S983, Y1008, S1039, S1063, S1136, S1159, T1225; for ENSP00000404047, sites are *T425*, Y459, T586, Y601, S722, Y747, S775 (S/Y for SNV), S778, S797, S807, S831, Y863, T890, S898, T964, Y972. NeXtProt lists 4 phosphotyrosine residues in 3 of the 4 ASVs while there are 7 N-linked glycan structures reported for the major ASV but only 1 structure for other ASVs. To denote an amino acid change in an ASV relative to ENSP00000269571 we italicize the residue notation and in the case of an altered site we have underlined the residue. In the cases where there are lesser numbers of PTMs listed for lower abundance ASVs this observation may be due to lack of experimental evidence rather than absence of PTMs. At this stage of characterization of the proteogenome there is little experimental evidence for PTMs in protein variants produced by misssense SNPs but the polymorphism of serine to tyrosine (residue 1051 in ENSP269571) would be expected to affect that site of modification.

The C-HPP initiative believes that the management of such a complex data set as illustrated here (and with regular updates) is best served by an informatics system and associated interfaces that can integrate such information from a diversity of information sources. This new type of informatics system will be the subject of a separate report in this issue[27].

Section 7 Cancer associated studies

We have chosen as an example the important oncogene, ERBB2(HER2), which resides on chromosome 17 and illustrates the value of considering proteomic observations in the context of the environment of the chromosome region in which the corresponding gene resides. ERBB2 encodes for a 185 kDa transmembrane glycoprotein that belongs to the family of epidermal growth factor receptors (EGFRs). Other members of this family include EGFR (ERBB1/HER1), ERBB3 (HER3), and ERBB4 (HER4). Ligand binding to the extracellular domain of these receptors induces homodimer (e.g. EGFR–EGFR) or heterodimer (e.g. EGFR–ERBB2) formation leading to the activation of the intracellular tyrosine kinase domain and subsequent signaling cascade[28]. Several ligands capable of activating the ERBB receptors have been identified, EGF-like ligands binding to EGFR and neuregulins (NRG1 (ERBB2), NRG2 (ERBB3, 4), NRG3 and 4 (ERBB4)[29,30]. ERBB2 has been shown to be a preferable interaction partner for all other ERBB receptors and such heterodimers are long lived and have a particularly high signaling potency[31,32].

The term amplicon is used to define gene amplification or the selective increase in the copy number of an oncogene and adjacent genes that can occur in development of solid tumors and should not be confused with elevated gene expression[33]. In breast and other cancers oncogene amplification can span several megabases and has been observed by comparative genome hybridization (CGH) on chromosomes 1q, 8p12, 8q24, 11q13, 12q13, 17q21, 17q23, and 20q13[28]. High resolution analysis of somatic copy number alterations from over 3000 cancer specimens identified 76 regions of significant amplifications and 25 regions which contained oncogenes such as MYC, CCND1, ERBB2, CDK4, NKX2-1, MDM2, EGFR, FGFR1 and KRAS[34, 35].

The ERBB2 amplicon observed in breast and other cancers, e.g. gastric, cervical, ovarian, pancreatic, and prostate, occurs typically over a 1.5Mb region and covers multiple genes in the region 17q12-q21 of chromosome 17[36,37,40]. The amplicon may be observed as homogeneously staining regions, or extrachromosomally, as cytogenetically visible double minute chromosomes or submicroscopic episomes[38]. The co-amplified genes may have an impact on the phenotype and clinical characteristics of ERBB2-amplified tumors as co-amplified genes may contribute to disease progression[40]. Genes in close proximity to ERBB2 and observed in tumor studies include the following: TIAF1, TRAF4, PSMB3, LASP1, MED1, CDK12, PPP1R1B, STARD3, PNMT, PGAP3, ERBB2, MIEN1, GRB7, IKZF3, ORMDL3, GSDMB, PSMD3, MED24, NR1D1, CASC3, CDC6, RARA, TOP2A, listed in order from the centromeric and telomeric ends of the amplicon (17q11.2 to q21.2). The amplicon can range from the minimal set of ERBB2 - GRB7 to the genes listed above[38, 39]. Moreover, increased expression of this set of genes was directly linked to gene amplification via copy number analysis of ERBB2 and adjacent genes. However, not all genes in this region are amplified in tumor samples and thus not included in the above list, for example the expression levels of PNMT were generally rather low, while expression of TCAP and NEUROD2 was either very weak or absent in these tumor samples[42–44].

The existence of the ERBB2 amplicon is a strong example of the value of the integration of proteomics with transcriptomic data with the potential for discovery of additional protein features of disease interest in a given data set. In Fig.6 we plot the data for two breast cancer cell lines that express high levels of ERBB2, namely SKBR3 and SUM190 (breast cancer). The figure shows the value of integrating RNASeq measurement with the corresponding proteomic data, e.g. PGAP3 MED1, CDK12, MED24 and CDC6 are not observed by proteomics but are part of the ERBB2 amplicon. Conversely the following members of the amplicon were observed by proteomics as well as RNASeq measurement: LASP1, ERBB2, MIEN1, GRB7, GSDMB, ORMDL3, PSMD3 and TOP2A which can give insights into the phenotypically important parts of the amplicon, namely actin cytoskeletal reorganization, regulation of apoptosis, stabilization of the phosphorylated form of ERBB2, endoplasmic reticulum-mediated Ca(+2) signaling and DNA synthesis.

In a follow up study we explored the ERBB2 amplicon in Oncomine[13], which contains the largest collection of curated cancer microarray data. Its integrated data-mining platform facilitates extensive meta-analyses across large numbers of mRNA datasets. We performed differential meta-analyses on human breast cancer microarray data sets and studied the commonly over-expressed chromosome 17 genes from ERBB2 positive, Estrogen receptor positive and Triple negative (ERBB2/ER/PR-negative) breast cancer subtypes. The top 20 over-expressed genes in each of the 3 breast cancer subtypes are given in Table 3. Remarkably we found no overlap between the top 20 genes expressed in these 3 different major breast cancer types.

In a separate analysis, we identified the 20 most over-expressed genes from all chromosomes across the ten ERBB2+ breast cancer microarray datasets in Oncomine. Remarkably, 13 of the top 20 genes were from chromosome 17 (Figure 7), with several clustered very near ERBB2. The genes (transcripts) are ranked by their association with ERBB2+ breast cancers; to compare these data with the consensus list of genes for the ERBB2 amplicon we have repeated the amplicon set identified in previous studies[30, 31] and underlined the genes overexpressed in Oncomine and from our cancer cell line studies (double underline denotes presence in both studies): TRAF4, TIAF1, PCGF2, PSMB3, LASP1, MED1, CDK12, PPP1R1B, STARD3, PNMT, PGAP3, ERBB2, MIEN1, GRB7, IKZF3, ORMDL3, GSDMB, PSMD3, MED24, NR1D1, CASC3, CDC6, TOP2A. While some genes in Oncomine are expressed at low levels in our study (e.g. TCAP, FBXL20) and are not included in this comparison, others are present on chromosome 17 but outside the amplicon e.g. CYB561 (see Table 3). It is noteworthy that most of the genes in close proximity to ERBB2 are highly ranked in Oncomine and observed in our studies. Thus the combined view of experimental data generated in our laboratories together with curated literature information strengthen the case for the integration of trancriptomic and proteomic data in the study of this amplicon and will be explored further with studies of ERBB2+ breast, gastric and colon cancer patient samples. Such a comparison is particularly germane since there is clinical evidence that Herceptin does not seem to yield benefit in ERBB2+ colorectal cancer patients and may have mixed results in such gastric cancer patients[47].

Concluding Remarks

We have demonstrated an organized approach to defining the baseline of what is currently known about protein products of the protein-coding genes on Chromosome 17, utilizing and comparing multiple valuable data resources. Post-translational modifications, sequence polymorphisms, and splice variants have been documented, features which must be studied at the protein level and to this end we have integrated of transcriptomics, mass spectrometry and antibody protein capture approaches, which is a model for integrated analyses of additional platforms like epigenomic and metabolomics data. We have highlighted the cancer-associated genes on this chromosome and especially the genes associated with ERBB2 (HER2/NEU); the striking regulation and co-expression of genes located in the ERBB2 amplicon from 17q12 to 17q23 demonstrates the value of a chromosome-centric approach to even such a complex phenotype as a particular subtype of human breast cancers.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We wish to thank the following for valuable discussions: Dan Rhodes, Compendia Bioscience, Inc, Ann Arbor, Michigan (Oncomine analysis) and Mark Fink for the insight that PGAP was part of the ERBB2 amplicon. This work was supported by following research grants: (Korea) The World Class University program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10086-0 (W.S.H. and Y.-K.P.), National Project for the Personalized Genomic Medicine A111218-11-CP01 (to YKP) from the Korean Ministry of Health and Welfare: (USA) The National Institutes of Health grants, U01-CA128427 to WSH, U54DA021519, UL1 RR024986, RM-08-029, and U54ES017885 to G.S.O.; R41 GM103362 to DF, NIH grant (MPS and H.I.); Texas State Rider for the Morgan Welch Inflammatory Breast Cancer Program and the G. Morris Dorrance Jr. Chair in Medical Oncology (M.C., Z.M.); NHGRI grants R01 GM087221, 2P50 GM076547 (E.D.): (Europe) EU FP7 grant 'ProteomeXchange' [grant number 260558], and from the Luxembourg Centre for Systems Biomedicine and the University of Luxembourg (E.W.D.); SIB Swiss Institute of Bioinformatics; Genebio SA; the Swiss Confederation's Commission for Technology and Innovation (CTI, grant 10214.1 PFLS-LS). Eurostars grant 6715 (A.B., P.G.); (Sweden) the Knut and Alice Wallenberg foundation for funding the HPA project (E.L, M.U.)

Reference

1. Paik YK, Jeong SK, Omenn GS, Uhlen M, Hanash S, Cho SY, Lee HJ, Na K, Choi EY, Yan FF, Zhang F, Zhang Y, Snyder M, Cheng Y, Chen R, Marko-Varga G, Deutsch EW, Kim H, Kwon JY, Aebersold R, Bairoch A, Taylor AD, Kim KY, Lee EY, Hochstrasser D, Legrain P, Hancock WS. The Chromosome-Centric Human Proteome Project for cataloging proteins encoded in the genome. Nat Biotechnol. 2012; 30(3):221–223. [PubMed: 22398612]
2. Hancock W, Omenn G, Legrain P, Paik YK. Proteomics, Human Proteome Project, and Chromosomes. J Proteome Res. 2011; 10(1):210–210. [PubMed: 21114295]
3. Zody MC, Garber M, Adams DJ, Sharpe T, Harrow J, Lupski JR, Nicholson C, Searle SM, Wilming L, Young SK, Abouelleil A, Allen NR, Bi W, Bloom T, Borowsky ML, Bugalter BE, Butler J, Chang JL, Chen CK, Cook A, Corum B, Cuomo CA, de Jong PJ, DeCaprio D, Dewar K, FitzGerald M, Gilbert J, Gibson R, Gnerre S, Goldstein S, Grafham DV, Grocock R, Hafez N, Hagopian DS, Hart E, Norman CH, Humphray S, Jaffe DB, Jones M, Kamal M, Khodiyar VK, LaButti K, Laird G, Lehoczky J, Liu X, Lokyitsang T, Loveland J, Lui A, Macdonald P, Major JE, Matthews L, Mauceli E, McCarroll SA, Mihalev AH, Mudge J, Nguyen C, Nicol R, O'Leary SB, Osoegawa K, Schwartz DC, Shaw-Smith C, Stankiewicz P, Steward C, Swarbreck D, Venkataraman V, Whittaker CA, Yang X, Zimmer AR, Bradley A, Hubbard T, Birren BW, Rogers J, Lander ES, Nusbaum C. DNA sequence of human chromosome 17 and analysis of rearrangement in the human lineage. Nature. 2006 Apr 20; 440(7087):1045–1049. [PubMed: 16625196]
4. Semon M, Duret L. Evolutionary origin and maintenance of coexpressed gene clusters in mammals. Mol Biol Evol. 2006; 23(9):1715–1723. [PubMed: 16757654]
5. Orsetti B, Nugoli M, Cervera N, Lasorsa L, Chuchana P, Ursule L, Nguyen C, Redon R, du Manoir S, Rodriguez C, Theillet C. Genomic and expression profiling of chromosome 17 in breast cancer reveals complex patterns of alterations and novel candidate genes. Cancer Res. 2004; 64(18):6453–6460. [PubMed: 15374954]
6. Zhou Y, Luoh SM, Zhang Y, Watanabe C, Wu TD, Ostland M, Wood WI, Zhang ZM. Genome-wide identification of chromosomal regions of increased tumor expression by transcriptome analysis. Cancer Res. 2003; 63(18):5781–5784. [PubMed: 14522899]

7. Kallioniemi A, Kallioniemi OP, Piper J, Tanner M, Stokke T, Chen L, Smith HS, Pinkel D, Gray JW, Waldman FM. Detection and Mapping of Amplified DNA-Sequences in Breast-Cancer by Comparative Genomic Hybridization. P Natl Acad Sci USA. 1994; 91(6):2156–2160.

8. Sproul D, Gilbert N, Bickmore WA. The role of chromatin structure in regulating the expression of clustered genes. Nat Rev Genet. 2005; 6(10):775–781. [PubMed: 16160692]

9. Lane L, Argoud-Puy G, Britan A, Cusin I, Duek PD, Evalet O, Gateau A, Gaudet P, Gleizes A, Masselot A, Zwahlen C, Bairoch A. neXtProt: a knowledge platform for human proteins. Nucleic Acids Res. 2012; 40(D1):D76–D83. [PubMed: 22139911]

10. The UniProt Consortium; Reorganizing the protein space at the Universal Protein Resource (UniProt), Nucleic Acids Res. 2012; 40(D1):D71–D75. m. [PubMed: 21447597]

11. Deutsch EW. The PeptideAtlas Project. Methods Mol Biol. 2010; 604:285–296. [PubMed: 20013378]

12. Zhang H, Loriaux P, Eng J, Campbell D, Keller A, Moss P, Bonneau R, Zhang N, Zhou Y, Wollscheid B, Cooke K, Yi EC, Lee H, Peskind ER, Zhang J, Smith RD, Aebersold R. UniPep--a database for human N-linked glycosites: a resource for biomarker discovery. Genome Biol. 2006; 7(8):R73. [PubMed: 16901351]

13. Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, Chinnaiyan AM. ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia. 2004; 6(1):1–6. [PubMed: 15068665]

14. Craig R, Cortens JP, Beavis RC. Open source system for analyzing, validating, and storing protein identification data. J Proteome Res. 2004; 3(6):1234–1242. [PubMed: 15595733]

15. Safran M, Dalah I, Alexander J, Rosen N, Iny Stein T, Shmoish M, Nativ N, Bahir I, Doniger T, Krug H, Sirota-Madi A, Olender T, Golan Y, Stelzer G, Harel A, Lancet D. GeneCards Version 3: the human gene integrator. Database (Oxford). 2010 baq020.

16. Desiere F, Deutsch EW, King NL, Nesvizhskii AI, Mallick P, Eng J, Chen S, Eddes J, Loevenich SN, Aebersold R. The PeptideAtlas project. Nucleic Acids Res. 2006; 34:D655–D658. [PubMed: 16381952]

17. Chen R, Mias GI, Li-Pook-Than J, Jiang LH, Lam HYK, Chen R, Miriami E, Karczewski KJ, Hariharan M, Dewey FE, Cheng Y, Clark MJ, Im H, Habegger L, Balasubramanian S, O'Huallachain M, Dudley JT, Hillenmeyer S, Haraksingh R, Sharon D, Euskirchen G, Lacroute P, Bettinger K, Boyle AP, Kasowski M, Grubert F, Seki S, Garcia M, Whirl-Carrillo M, Gallardo M, Blasco MA, Greenberg PL, Snyder P, Klein TE, Altman RB, Butte AJ, Ashley EA, Gerstein M, Nadeau KC, Tang H, Snyder M. Personal Omics Profiling Reveals Dynamic Molecular and Medical Phenotypes. Cell. 2012; 148(6):1293–1307. [PubMed: 22424236]

18. Geiger T, Wehner A, Schaab C, Cox J, Mann M. Comparative Proteomic Analysis of Eleven Common Cell Lines Reveals Ubiquitous but Varying Expression of Most Proteins. Mol Cell Proteomics. 2012; 11(3)

19. Strachan, T.; Read, AP. Human Molecular Genetics. New York: Wiley-Liss; 1999. Chapter 7.2.

20. Grouse L, Omenn GA, McCarthy BJ. Studies by DNA-RNA hybridization of transcriptional diversity in human brain. Journal of neurochemistry. 1973; 20(4):1063–1073. [PubMed: 4697869]

21. Revil T, Toutant J, Shkreta L, Garneau D, Cloutier P, Chabot B. Protein kinase C-dependent control of Bcl-x alternative splicing. Mol Cell Biol. 2007; 27(24):8431–8441. [PubMed: 17923691]

22. Wegran F, Boidot R, Oudin C, Riedinger JM, Bonnetain F, Lizard-Nacol S. Overexpression of caspase-3s splice variant in locally advanced breast carcinoma is associated with poor response to neoadjuvant chemotherapy. Clin Cancer Res. 2006; 12(19):5794–5800. [PubMed: 17020986]

23. Yu DH, Hung MC. Overexpression of ErbB2 in cancer and ErbB2-targeting strategies. Oncogene. 2000; 19(53):6115–6121. [PubMed: 11156524]

24. Connolly D, Yang ZX, Castaldi M, Simmons N, Oktay MH, Coniglio S, Fazzari MJ, Verdier-Pinard P, Montagna C. Septin 9 isoform expression, localization and epigenetic changes during human and mouse breast cancer progression. Breast Cancer Res. 2011; 13(4)

25. Lee A, Kolarich D, Haynes PA, Jensen PH, Baker MS, Packer NH. Rat liver membrane glycoproteome: enrichment by phase partitioning and glycoprotein capture. J Proteome Res. 2009; 8(2):770–781. [PubMed: 19125615]

26. Build for Human. http://www.ncbi.nlm.nih.gov/projects/SNP

27. Goode, Robert JA.; Yu, Simon; Kannan, Anitha; Christiansen, Jeffrey H.; Beitz, Anthony; Hancock, William S.; Smith, A. Ian; Nice, Edouard C. The Proteome Browser web portal. J.Prot.Res. Manuscript submitted.

28. Rubin I, Yarden Y. The basic biology of HER2. Ann Oncol. 2001; 12:3–8.

29. PinkasKramarski R, Eilam R, Alroy I, Levkowitz G, Lonai P, Yarden Y. Differential expression of NDF/neuregulin receptors ErbB-3 and ErbB-4 and involvement in inhibition of neuronal differentiation. Oncogene. 1997; 15(23):2803–2815. [PubMed: 9419971]

30. Riese DJ, Stern DF. Specificity within the EGF family ErbB receptor family signaling network. Bioessays. 1998; 20(1):41–48. [PubMed: 9504046]

31. GrausPorta D, Beerli RR, Daly JM, Hynes NE. ErbB-2, the preferred heterodimerization partner of all ErbB receptors, is a mediator of lateral signaling. Embo J. 1997; 16(7):1647–1655. [PubMed: 9130710]

32. Tzahar E, Waterman H, Chen XM, Levkowitz G, Karunagaran D, Lavi S, Ratzkin BJ, Yarden Y. A hierarchical network of interreceptor interactions determines signal transduction by neu differentiation factor/neuregulin and epidermal growth factor. Mol Cell Biol. 1996; 16(10):5276–5287. [PubMed: 8816440]

33. Schwab M. Oncogene amplification in solid tumors. Semin Cancer Biol. 1999; 9(4):319–325. 34. [PubMed: 10448118] Myllykangas S, Knuutila S. Manifestation, mechanisms and mysteries of gene amplifications. Cancer Lett. 2006; 232(1):79–89. [PubMed: 16288831]

34. Yao J, Weremowicz S, Feng B, Gentleman RC, Marks JR, Gelman R, Brennan C, Polyak K. Combined cDNA array comparative genomic hybridization and serial analysis of gene expression analysis of breast tumor progression. Cancer Res. 2006; 66(8):4065–4078. [PubMed: 16618726]

35. Kauraniemi P, Barlund M, Monni O, Kallioniemi A. New amplified and highly expressed genes discovered in the ERBB2 amplicon in breast cancer by cDNA microarrays. Cancer Res. 2001; 61(22):8235–8240. [PubMed: 11719455]

36. Beroukhim R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M, Mc Henry KT, Pinchback RM, Ligon AH, Cho YJ, Haery L, Greulich H, Reich M, Winckler W, Lawrence MS, Weir BA, Tanaka KE, Chiang DY, Bass AJ, Loo A, Hoffman C, Prensner J, Liefeld T, Gao Q, Yecies D, Signoretti S, Maher E, Kaye FJ, Sasaki H, Tepper JE, Fletcher JA, Tabernero J, Baselga J, Tsao MS, Demichelis F, Rubin MA, Janne PA, Daly MJ, Nucera C, Levine RL, Ebert BL, Gabriel S, Rustgi AK, Antonescu CR, Ladanyi M, Letai A, Garraway LA, Loda M, Beer DG, True LD, Okamoto A, Pomeroy SL, Singer S, Golub TR, Lander ES, Getz G, Sellers WR, Meyerson M. The landscape of somatic copy-number alteration across human cancers. Nature. 2010; 463(7283):899–905. [PubMed: 20164920]

37. Dressman MA, Baras A, Malinowski R, Alvis LB, Kwon I, Walz TM, Polymeropoulos MH. Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer. Cancer Res. 2003; 63(9):2194–2199. [PubMed: 12727839]

38. Dressman MA, Baras A, Malinowski R, Alvis LB, Kwon I, Walz TM, Polymeropoulos MH. Gene expression profiling detects gene amplification and differentiates tumor types in breast cancer. Cancer Res. 2003; 63(9):2194–2199. [PubMed: 12727839]

39. Katoh M, Katoh M. Evolutionary recombination hotspot around GSDML-GSDM locus is closely linked to the oncogenomic recombination hotspot around the PPP1R1B-ERBB2-GRB7 amplicon. Int J Oncol. 2004; 24(4):757–763. [PubMed: 15010812]

40. Kauraniemi P, Kallioniemi A. Activation of multiple cancer-associated genes at the ERBB2 amplicon in breast cancer. Endocr-Relat Cancer. 2006; 13(1):39–49. [PubMed: 16601278]

41. Sircoulomb F, Bekhouche I, Finetti P, Adelaide J, Ben Hamida A, Bonansea J, Raynaud S, Innocenti C, Charafe-Jauffret E, Tarpin C, Ben Ayed F, Viens P, Jacquemier J, Bertucci F, Birnbaum D, Chaffanet M. Genome profiling of ERBB2-amplified breast cancers. Bmc Cancer. 2010; 10:539. [PubMed: 20932292]

42. Luoh SW. Amplification and expression of genes from the 17q11 similar to q12 amplicon in breast cancer cells. Cancer Genet Cytogen. 2002; 136(1):43–47.

43. Tomasetto C, Regnier C, Mooglutz C, Mattei MG, Chenard MP, Lidereau R, Basset P, Rio MC. Identification of 4 Novel Human Genes Amplified and Overexpressed in Breast-Carcinoma and

Localized to the Q11-Q21.3 Region of Chromosome-17. Genomics. 1995; 28(3):367–376. [PubMed: 7490069]

44. Bieche I, Tomasetto C, Regnier CH, MoogLutz C, Rio MC, Lidereau R. Two distinct amplified regions at 17q11-q21 involved in human primary breast cancer. Cancer Res. 1996; 56(17):3886–3890. [PubMed: 8752152]

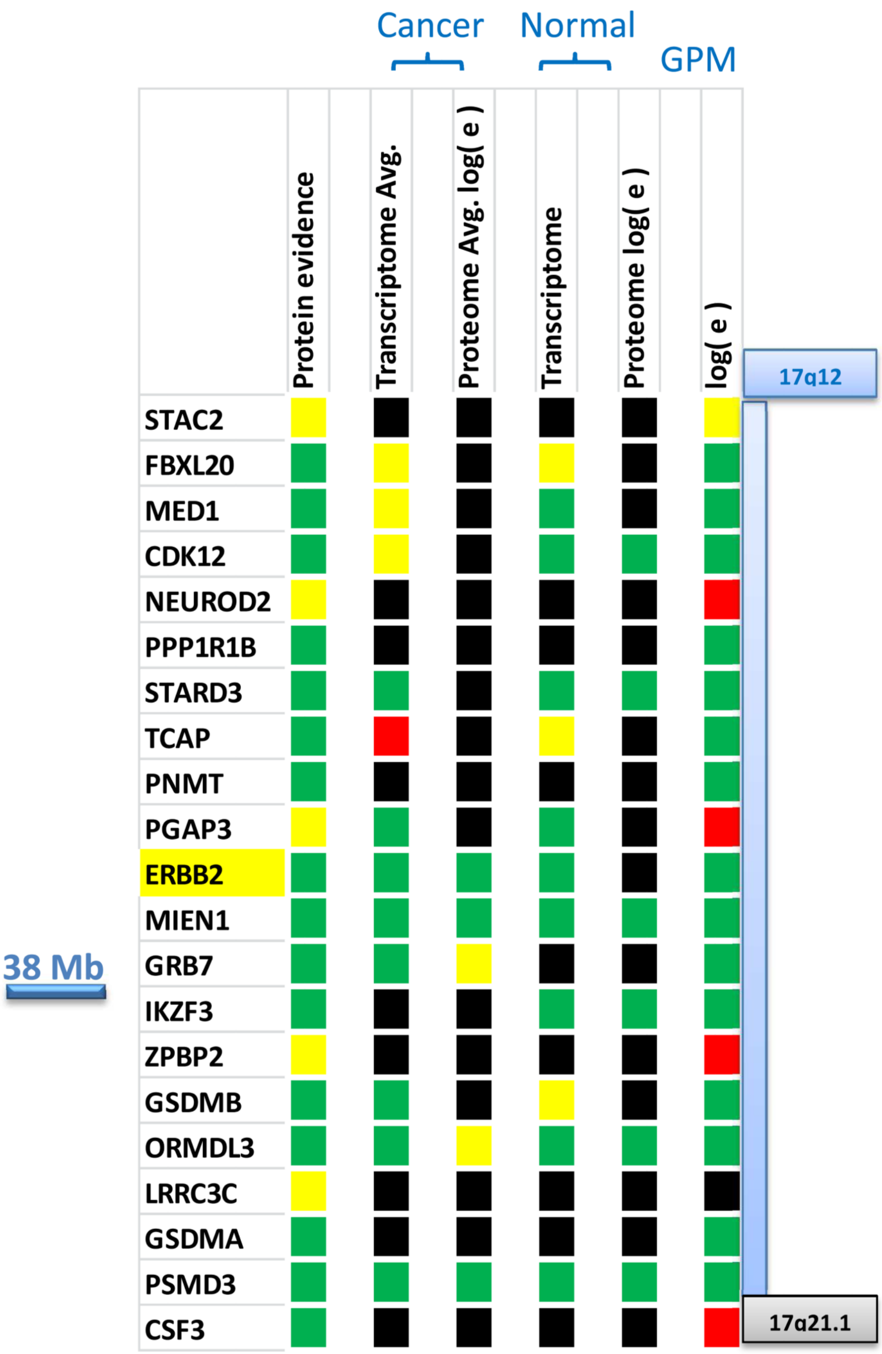
45. Schultz L, Khera S, Sleve D, Heath J, Chang NS. TIAF1 and p53 functionally interact in mediating apoptosis and silencing of TIAF1 abolishes nuclear translocation of serine 15-phosphorylated p53. DNA Cell Biol. 2004; 23(1):67–74. [PubMed: 14965474]

46. Almond JB, Cohen GM. The proteasome: a novel target for cancer chemotherapy. Leukemia. 2002; 16(4):433–443. [PubMed: 11960320]

47. Hasegawa N, Sumitomo A, Fujita A, Aritome N, Mizuta S, Matsui K, Ishino R, Inoue K, Urahama N, Nose J, Mukohara T, Kamoshida S, Roeder RG, Ito M. Mediator Subunits MED1 and MED24 Cooperatively Contribute to Pubertal Mammary Gland Development and Growth of Breast Carcinoma Cells. Mol Cell Biol. 2012; 32(8):1483–1495. [PubMed: 22331469]

48. Kourtidis A, Jain R, Carkner RD, Eifert C, Brosnan MJ, Conklin DS. An RNA Interference Screen Identifies Metabolic Regulators NR1D1 and PBP as Novel Survival Factors for Breast Cancer Cells with the ERBB2 Signature. Cancer Res. 2010; 70(5):1783–1792. [PubMed: 20160030]

49. Menendez JA, Vellon L, Lupu R. DNA topoisomerase II alpha (TOP2A) inhibitors up-regulate fatty acid synthase gene expression in SK-Br3 breast cancer cells: In vitro evidence for a 'functional amplicon' involving FAS, Her-2/neu and TOP2A genes. Int J Mol Med. 2006; 18(6):1081–1087. [PubMed: 17089011]



		Cancer		Normal		GPM
	Protein evidence	Transcriptome Avg.	Proteome Avg. log(e)	Transcriptome	Proteome log(e)	log(e)
STAC2	🟨	⬛	⬛	⬛	⬛	🟨
FBXL20	🟩	🟨	⬛	🟨	⬛	🟩
MED1	🟩	🟨	⬛	🟩	⬛	🟩
CDK12	🟩	🟨	⬛	🟩	🟩	🟩
NEUROD2	🟨	⬛	⬛	⬛	⬛	🟥
PPP1R1B	🟩	⬛	⬛	⬛	⬛	🟩
STARD3	🟩	🟩	⬛	🟩	🟩	🟩
TCAP	🟩	🟥	⬛	🟨	⬛	🟩
PNMT	🟩	⬛	⬛	⬛	⬛	🟩
PGAP3	🟨	🟩	⬛	🟩	⬛	🟥
ERBB2	🟩	🟩	🟩	🟩	⬛	🟩
MIEN1	🟩	🟩	🟩	🟩	🟩	🟩
GRB7	🟩	🟩	🟨	⬛	⬛	🟩
IKZF3	🟩	⬛	⬛	🟩	🟩	🟩
ZPBP2	🟨	⬛	⬛	⬛	⬛	🟥
GSDMB	🟩	🟩	⬛	🟨	⬛	🟩
ORMDL3	🟩	🟩	🟨	🟩	🟩	🟩
LRRC3C	🟨	⬛	⬛	⬛	⬛	⬛
GSDMA	🟩	⬛	⬛	⬛	⬛	🟩
PSMD3	🟩	🟩	🟩	🟩	🟩	🟩
CSF3	🟩	⬛	⬛	⬛	⬛	🟥

Figure 1. Parts list for the genomic region around ERBB2 (10 protein coding genes on each side)

The color coding used the following: green: protein level evidence in Uniprot, RNA-Seq RPKM >=15, proteomic data for the two studies and in GPMDB log(e) <-10; yellow: transcript level evidence, RNA-Seq RPKM 3~15, proteomics, log(e) −5~-10; red: uncertain protein evidence, RNA-Seq RPKM 1~3, proteomics log(e) −1~-5; black: no information. The cancer data were from two breast cancer cell lines: SKBR3 and SUM190 and normal sample consisted of peripheral mononuclear cells collected over serial time points from an individual. The samples were trypsinized and analyzed by nanoLC-MS/MS using a FTMS (cancer) or orbitrap (normal) linear ion-trap mass spectrometer. Strand-specific RNA-Seq libraries were prepared and sequenced on a lane of the Illumina HiSeq 2000 instrument per sample to obtain transcript data[15]. All RNA-Seq data are available at Short Read Archive (SRS366582, SRS366583, SRS366584, SRS366609, SRS366610, SRS366611). For proteomics data see: (http://gpmdb.thegpm.org/data/keyword/chr17%20hancock) and peripheral leukocyte data (http://gpmdb.thegpm.org/~/dblist_gpmnotes/keyword=pubmed%2022424236)

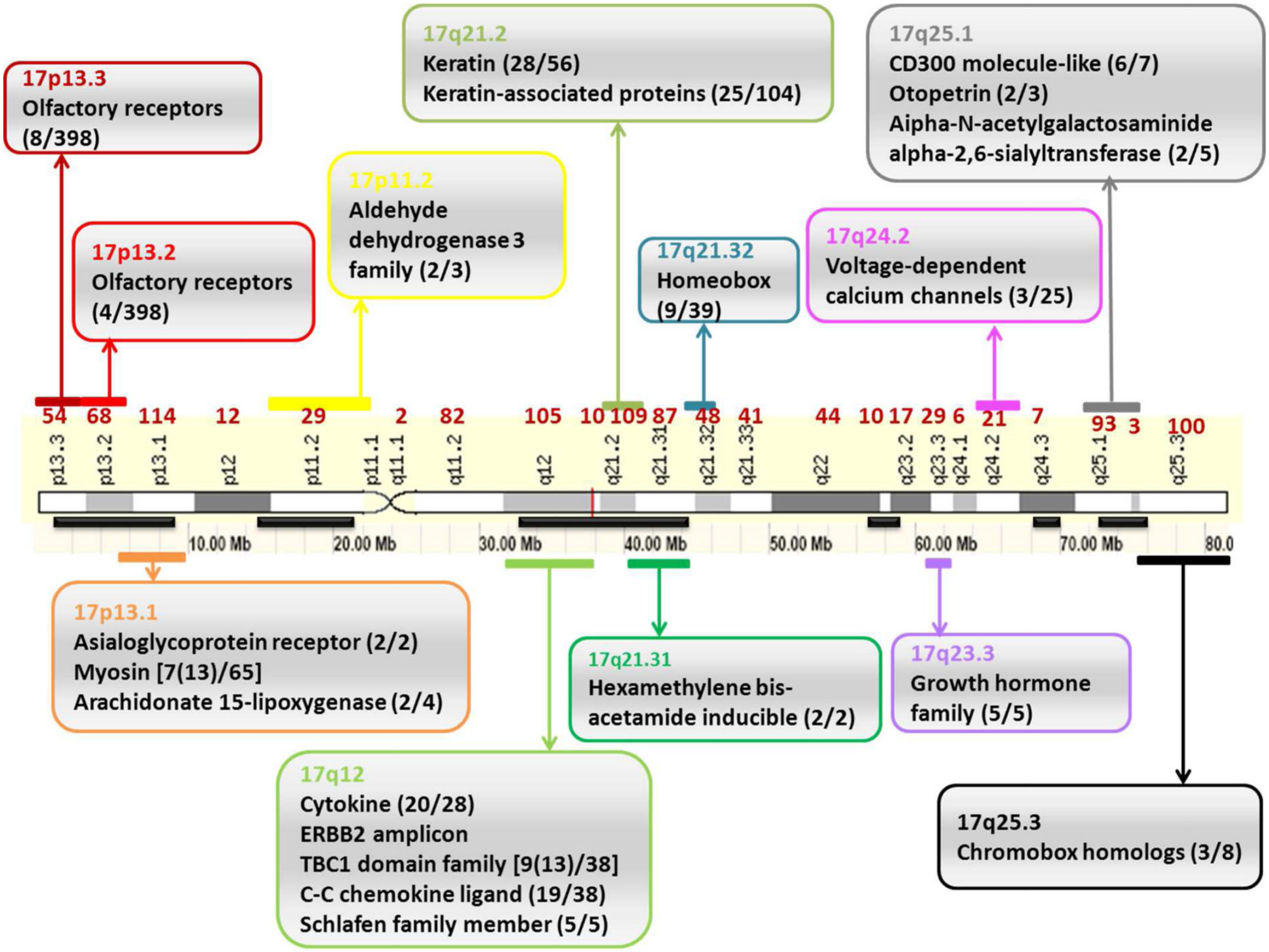
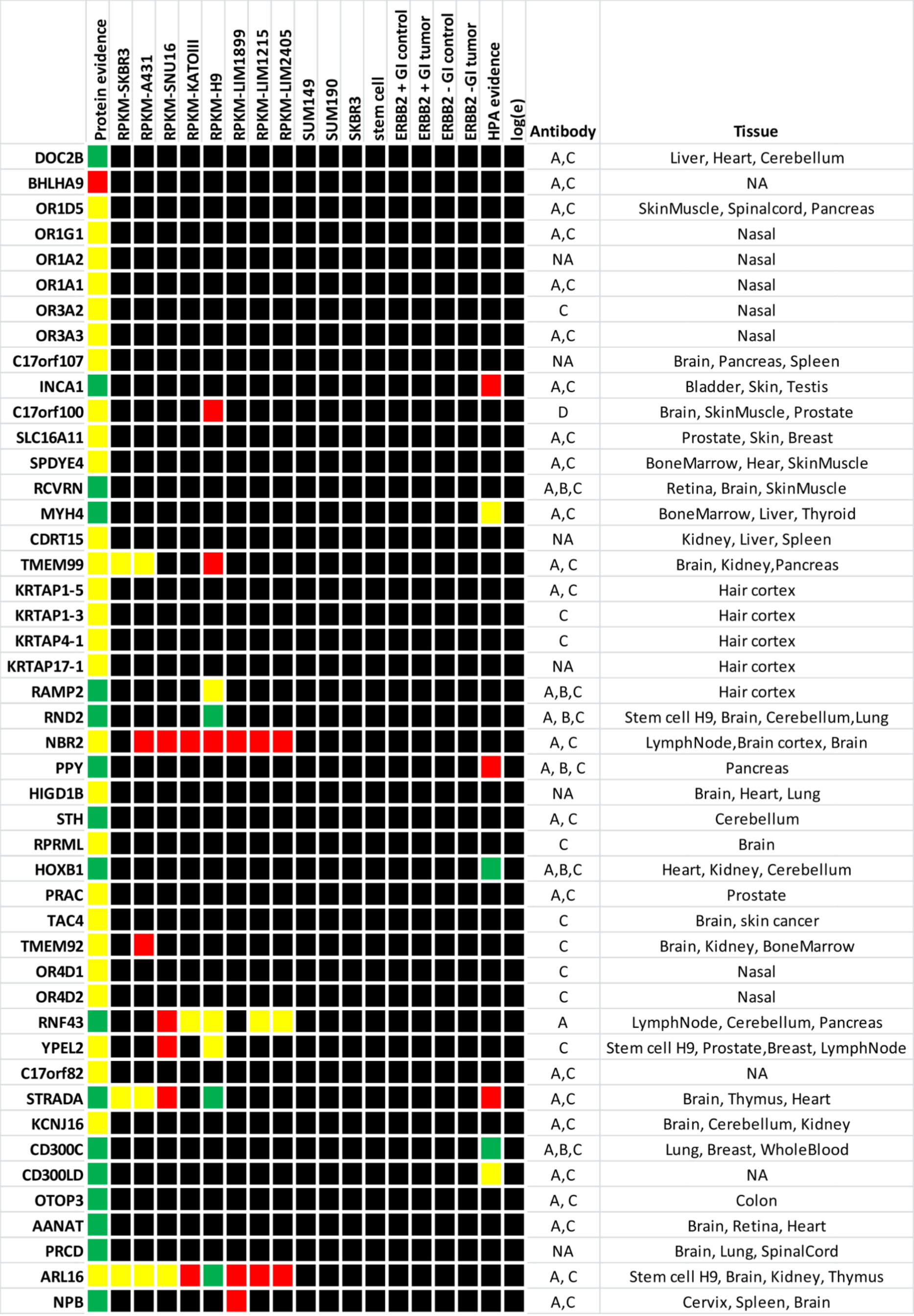

17p13.3
Olfactory receptors (8/398)
17q21.2
Keratin (28/56)
Keratin-associated proteins (25/104)
17q25.1
CD300 molecule-like (6/7)
Otopetrin (2/3)
Aipha-N-acetylgalactosaminide alpha-2,6-sialyltransferase (2/5)
17p11.2
Aldehyde dehydrogenase 3 family (2/3)
17p13.2
Olfactory receptors (4/398)
17q21.32
Homeobox (9/39)
17q24.2
Voltage-dependent calcium channels (3/25)
54
68
114
12
29
2
82
105
10
109
87
48
41
44
10
17
29
6
21
7
93
3
100
p13.3
p13.2
p13.1
p12
p11.2
p11.1
q11.1
q11.2
q12
q21.2
q21.31
q21.32
q21.33
q22
q23.2
q23.3
q24.1
q24.2
q24.3
q25.1
q25.3
10.00 Mb
20.00 Mb
30.00 Mb
40.00 Mb
50.00 Mb
60.00 Mb
70.00 Mb
80.0
17p13.1
Asialoglycoprotein receptor (2/2)
Myosin [7(13)/65]
Arachidonate 15-lipoxygenase (2/4)
17q21.31
Hexamethylene bis-acetamide inducible (2/2)
17q23.3
Growth hormone family (5/5)
17q12
Cytokine (20/28)
ERBB2 amplicon
TBC1 domain family [9(13)/38]
C-C chemokine ligand (19/38)
Schlafen family member (5/5)
17q25.3
Chromobox homologs (3/8)

Figure 2. Selected examples of protein families on chromosome 17

The protein families are shown in the boxes with band annotations. Within the brackets, the format is shown as [number of members of a given protein family in one region (total number of members on chromosome 17)/total family members in human genome]. The numbers of protein coding genes are listed above each band. The solid bars represent the regions with a gene density above the average on chromosome 17, namely 30.3 protein coding genes per Mb[15].



	Protein evidence	RPKM-SKBR3	RPKM-A431	RPKM-SNU16	RPKM-KATOIII	RPKM-H9	RPKM-LIM1899	RPKM-LIM1215	RPKM-LIM2405	SUM149	SUM190	SKBR3	stem cell	ERBB2 + GI control	ERBB2 + GI tumor	ERBB2 - GI control	ERBB2 -GI tumor	HPA evidence	log(e)	Antibody	Tissue
DOC2B	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Liver, Heart, Cerebellum
BHLHA9	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	NA
OR1D5	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	SkinMuscle, Spinalcord, Pancreas
OR1G1	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Nasal
OR1A2	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	NA	Nasal
OR1A1	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Nasal
OR3A2	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Nasal
OR3A3	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Nasal
C17orf107	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	NA	Brain, Pancreas, Spleen
INCA1	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟥	⬛	A,C	Bladder, Skin, Testis
C17orf100	🟨	⬛	⬛	⬛	⬛	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	D	Brain, SkinMuscle, Prostate
SLC16A11	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Prostate, Skin, Breast
SPDYE4	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	BoneMarrow, Hear, SkinMuscle
RCVRN	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,B,C	Retina, Brain, SkinMuscle
MYH4	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	⬛	A,C	BoneMarrow, Liver, Thyroid
CDRT15	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	NA	Kidney, Liver, Spleen
TMEM99	🟨	🟨	🟨	⬛	⬛	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A, C	Brain, Kidney,Pancreas
KRTAP1-5	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A, C	Hair cortex
KRTAP1-3	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Hair cortex
KRTAP4-1	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Hair cortex
KRTAP17-1	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	NA	Hair cortex
RAMP2	🟩	⬛	⬛	⬛	⬛	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,B,C	Hair cortex
RND2	🟩	⬛	⬛	⬛	⬛	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A, B,C	Stem cell H9, Brain, Cerebellum,Lung
NBR2	🟨	⬛	🟥	🟥	🟥	🟥	🟥	🟥	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A, C	LymphNode,Brain cortex, Brain
PPY	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟥	⬛	A, B, C	Pancreas
HIGD1B	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	NA	Brain, Heart, Lung
STH	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A, C	Cerebellum
RPRML	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Brain
HOXB1	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟩	⬛	A,B,C	Heart, Kidney, Cerebellum
PRAC	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Prostate
TAC4	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Brain, skin cancer
TMEM92	🟨	⬛	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Brain, Kidney, BoneMarrow
OR4D1	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Nasal
OR4D2	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Nasal
RNF43	🟩	⬛	⬛	🟥	🟨	🟨	⬛	🟨	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A	LymphNode, Cerebellum, Pancreas
YPEL2	🟨	⬛	⬛	🟥	⬛	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	C	Stem cell H9, Prostate,Breast, LymphNode
C17orf82	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	NA
STRADA	🟩	🟨	🟨	🟥	⬛	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟥	⬛	A,C	Brain, Thymus, Heart
KCNJ16	🟨	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Brain, Cerebellum, Kidney
CD300C	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟩	⬛	A,B,C	Lung, Breast, WholeBlood
CD300LD	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	🟨	⬛	A,C	NA
OTOP3	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A, C	Colon
AANAT	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Brain, Retina, Heart
PRCD	🟩	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	NA	Brain, Lung, SpinalCord
ARL16	🟨	🟨	🟨	🟨	🟥	🟩	🟥	🟥	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A, C	Stem cell H9, Brain, Kidney, Thymus
NPB	🟩	⬛	⬛	⬛	⬛	⬛	🟥	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	⬛	A,C	Cervix, Spleen, Brain

Figure 3. Uncharacterized gene products on chromosome 17

The gene names used are those listed in Ensembl. The following RNAseq data is illustrated (RPKM: reads per kilobase per million mapped reads): SKBR3, A431, SNU16, KATOIII, H9, LIM1899, LIM1215, LIM2405. The proteomics data arelisted: SUM149, SUM190, SKBR3, Stem cell, ERBB2+ GI control, ERBB2+ GI tumor, ERBB2- GI control, ERBB2- GI tumor. The color coding of HPA evidence are used as following rules: green (high), yellow (medium), low (red) and black (very low or NULL). The color coding of the rest of the table are the same as figure 1 (see figure 1 legend). Antibody information is shown as:

NA: not available, A: AntibodyPedia, polyclonal B: AntibodyPedia, monoclonal, C: Labome (http://www.antibodypedia.com, http://www.labome.com). The potential tissue sources information was obtained from GeneCards15.

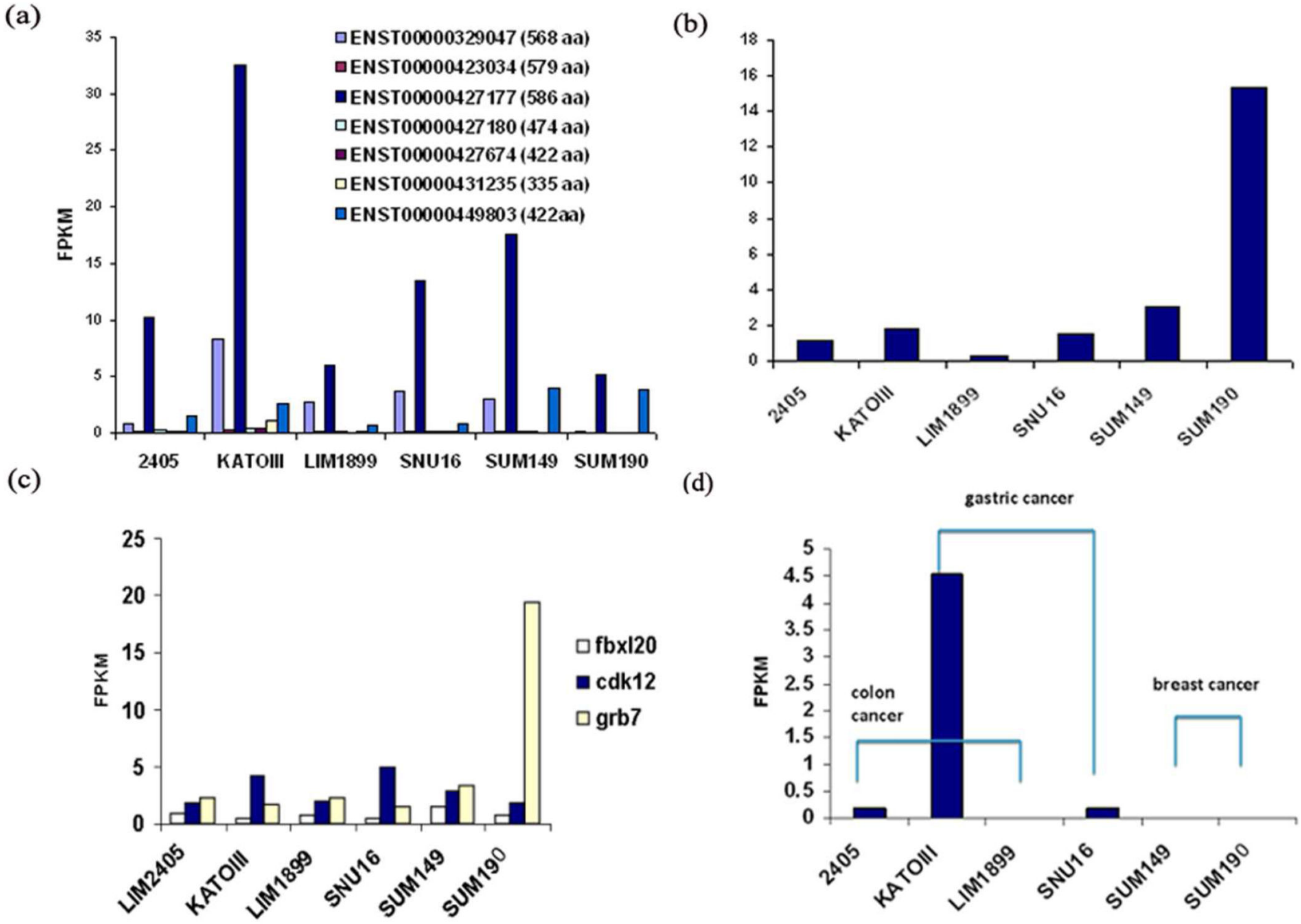


Figure 4.

(a) Septin 9 (SEPT9) transcript expression in the six ERBB2+ cell lines. (b) The relative abundance of the shorter variant of ERBB2 (ENSP00000385185) across the six cancer cell lines. (c) Relative expression levels of ASTs of three genes around ERBB2 on chromosome 17 in six ERBB2+ cell lines. Only the longest AST of each of these 3 genes was expressed. (d) Relative expression levels of the shorter transcript variant of Ppp1r1b in the six ErbB2 + cancer cell lines.

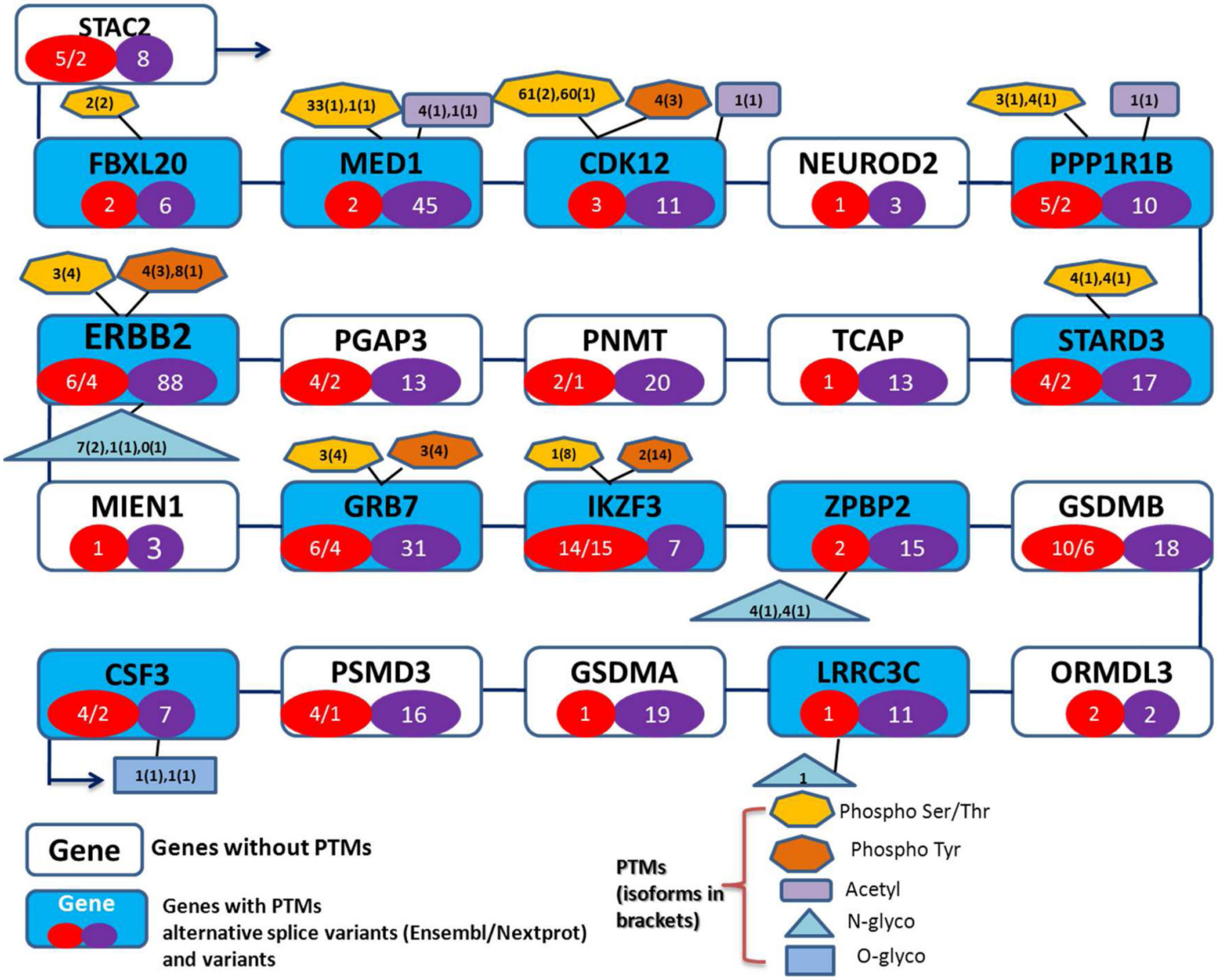
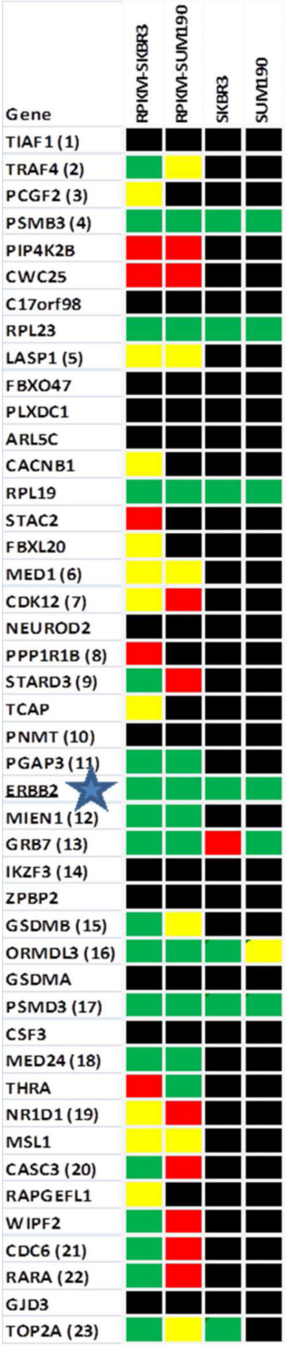


Figure 5. PTM information for genes around ERBB2

The following graphics are used to denote different types of information: Blue boxes have genes with PTM information; Red circles, number of alternative splice variants in Ensembl/neXtProt; Purple circles, number of variants in neXtProt; Hexagon shape in yellow, phosphorylation at Ser/Thr; Hexagon shape in orange, phosphorylation at Tyr; Rectangular shape acetylation; Triangle, N-glycosylation; square, O-glycosylation. The first number in each shape represents the number of PTMs for the major ASVs and a second number relates to the number of PTMs in secondary ASVs.



Gene	RPKM-SKBR3	RPKM-SUM190	SKBR3	SUM190
TIAF1 (1)	⬛	⬛	⬛	⬛
TRAF4 (2)	🟩	🟨	⬛	⬛
PCGF2 (3)	🟨	⬛	⬛	⬛
PSMB3 (4)	🟩	🟩	🟩	🟩
PIP4K2B	🟥	🟥	⬛	⬛
CWC25	🟥	🟥	⬛	⬛
C17orf98	⬛	⬛	⬛	⬛
RPL23	🟩	🟩	🟩	🟩
LASP1 (5)	🟨	🟨	⬛	⬛
FBXO47	⬛	⬛	⬛	⬛
PLXDC1	⬛	⬛	⬛	⬛
ARL5C	⬛	⬛	⬛	⬛
CACNB1	🟨	⬛	⬛	⬛
RPL19	🟩	🟩	🟩	🟩
STAC2	🟥	⬛	⬛	⬛
FBXL20	🟨	⬛	⬛	⬛
MED1 (6)	🟨	🟨	⬛	⬛
CDK12 (7)	🟨	🟥	⬛	⬛
NEUROD2	⬛	⬛	⬛	⬛
PPP1R1B (8)	🟥	⬛	⬛	⬛
STARD3 (9)	🟩	🟥	⬛	⬛
TCAP	🟨	⬛	⬛	⬛
PNMT (10)	⬛	⬛	⬛	⬛
PGAP3 (11)	🟩	🟩	⬛	⬛
ERBB2 ★	🟩	🟩	🟩	🟩
MIEN1 (12)	🟩	🟩	⬛	⬛
GRB7 (13)	🟩	🟩	🟥	🟩
IKZF3 (14)	⬛	⬛	⬛	⬛
ZPBP2	⬛	⬛	⬛	⬛
GSDMB (15)	🟩	🟨	⬛	⬛
ORMDL3 (16)	🟩	🟩	🟩	🟨
GSDMA	⬛	⬛	⬛	⬛
PSMD3 (17)	🟩	🟩	🟩	🟩
CSF3	⬛	⬛	⬛	⬛
MED24 (18)	🟩	🟩	⬛	⬛
THRA	🟥	🟩	⬛	⬛
NR1D1 (19)	🟨	🟥	⬛	⬛
MSL1	🟨	🟨	⬛	⬛
CASC3 (20)	🟩	🟥	⬛	⬛
RAPGEFL1	🟨	⬛	⬛	⬛
WIPF2	🟩	🟥	⬛	⬛
CDC6 (21)	🟩	🟥	⬛	⬛
RARA (22)	🟩	🟥	⬛	⬛
GJD3	⬛	⬛	⬛	⬛
TOP2A (23)	🟩	🟨	🟩	⬛

Figure 6. The set of protein coding genes comprising the ERBB2 amplicon which presents transcriptomic and proteomic information determined for ERBB2-expressing breast cancer cell lines (SKBR3 and SUM190)

The color coding is listed in figure 1 legend. ERBB2 is denoted with a star. The presence of a gene with cancer associations is denoted with a number and the following information, which was collected from GeneCards as well as cited references.

1. TIAF1, anti-apoptotic factor, induced by TGFB1, functionally interacts with p53 in regulating apoptosis[45].

2. TRAF4, commonly overexpressed in a wide range of tumors, adaptor protein and signal transducer, links members of the tumor necrosis factor receptor (TNFR) family to signaling pathways, regulation of apoptosis.
3. PCGF2, transcriptional and tumor suppressor, a diagnostic marker for poor prognosis in breast and prostate cancer patients.
4. PSMB3, proteasome subunit, beta type, 3, ubiquitin-proteasome system is an important regulator of cell growth and apoptosis[46].
5. LASP1, regulation of dynamic actin-based, cytoskeletal activities and zyxin localization in breast carcinomas, tumor growth and migration in cancer, enhances proliferation of ovarian and colorectal cancer.
6. MED1 or PPAR binding protein (PPARBP), nuclear co-activator, activates the transcription of vitamin D receptor, retinoid acid receptor and estrogen receptor, involved in cell growth, differentiation and amplified in a subset of breast tumors, regulates p53-dependent apoptosis.
7. CDK12, deletions within 17q12 region leading to CDK12-ERBB2 fusion protein, related to gastric cancer.
8. PPP1R1B, (protein phosphatase 1, regulatory inhibitor, subunit 1B) signaling member of wnt pathways that is frequently overexpressed in breast, prostate, colon, and stomach carcinomas.
9. STARD3, overexpression in cancer cells increases steroid hormone production, promoting growth of hormone-responsive tumors such as breast cancer.
10. PNMT, phenylethanolamine N-methyltransferase which converts noradrenaline to adrenaline, observed by increased gene copy number in breast cancer.
11. PGAP3, lipid remodeling steps of GPI-anchor maturation, key step in lipid raft assembly of ERBB2 heterodimers.
12. MIEN1, migration and invasion enhancer, overexpression in various breast and prostate cancer, enhances migration and invasion of tumor cells, regulation of apoptosis.
13. GRB7, growth factor receptor-bound protein 7 binds to tyrosine phosphorylated HER2, promotes activation of HRAS, associated with invasive breast, ovarian, gastric prostate and esophageal carcinomas.
14. IKZF3, IKAROS family zinc finger 3, transcription factor, major tumor suppressor involved in human B-cell acute lymphoblastic leukemia, interacts with HRAS.
15. Gasdermin-like (GSDML), linked to cancer development and progression, involved in secretory pathways.
16. ORMDL3, negative regulator of sphingolipid synthesis, may indirectly regulate endoplasmic reticulum-mediated Ca(+2) signaling
17. PSMD3, proteasome 26S subunit, non-ATPase, 3 (P58), ubiquitin-proteasome system is an important regulator of cell growth and apoptosis[46].

18. MED24, component of the Mediator complex, a coactivator of RNA polymerase II-dependent genes, mediates growth of breast carcinoma cells[47].

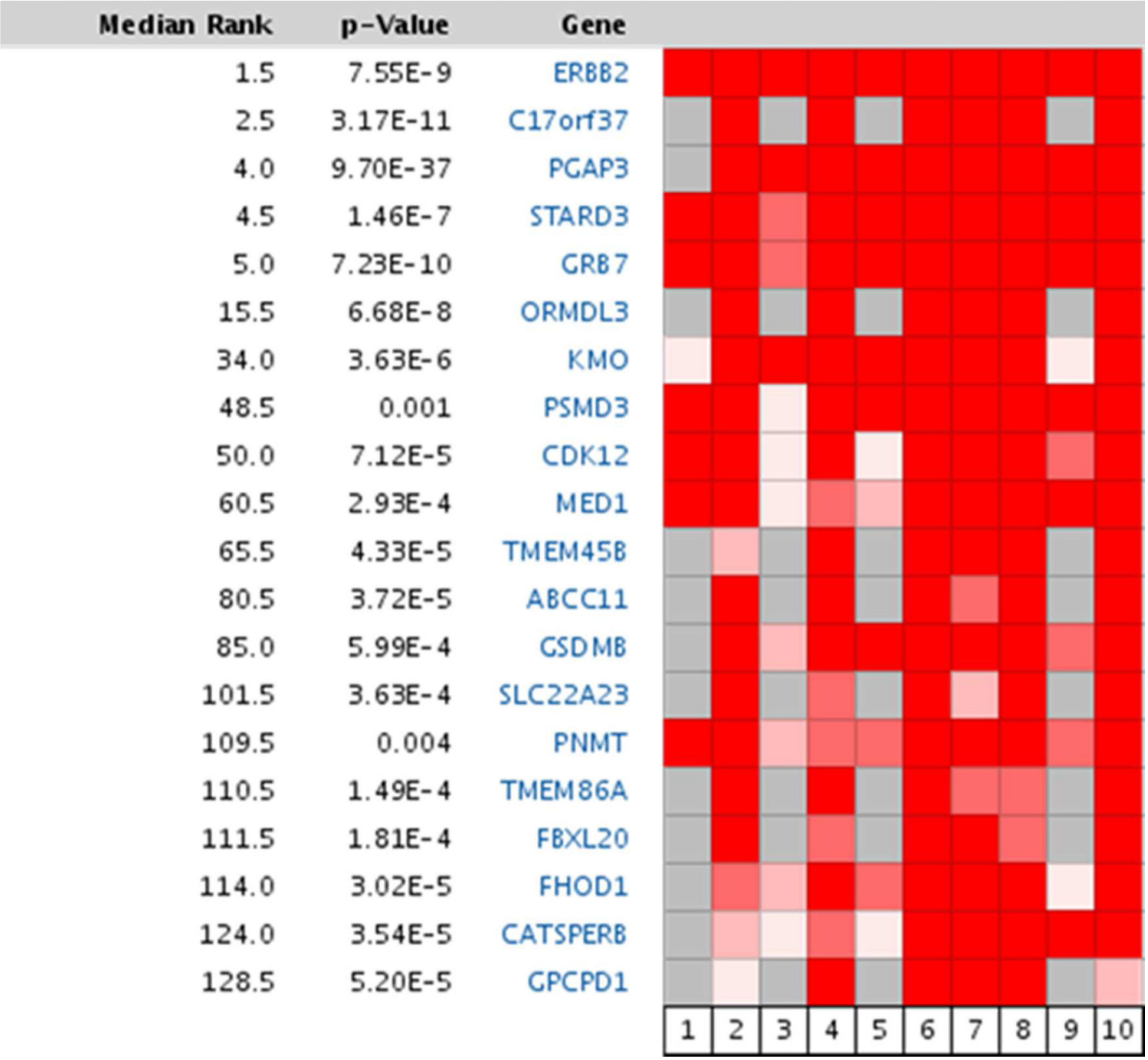
19. NR1D1, nuclear receptor subfamily 1, group D, required for energy production in ERBB2 expressing breast cancer cells[48].

20. CASC3, cancer susceptibility candidate 3, component of mRNA splicing-dependent multi-protein exon junction complex (EJC), overexpressed in breast and gastric cancer.

21. CDC6, cell division cycle 6 homolog, regulator at the early steps of DNA replication, transcription regulated by mitogenic signals, overexpressed or associated with prostate, squamous, cervical, lung and liver cancer.

22. RARA, retinoic acid receptor, alpha, mediates embryogenesis, differentiation and growth arrest, some ER-negative breast cancer cell lines (SKBR3) express high levels of RAR alpha protein and RARA has been observed as part of the ERBB2 amplicon.

23. TOP2A, topoisomerase IIa, markedly upregulated in breast, prostate, gastric, ovarian and lung cancer, catalyzing the ATP dependent breakage and rejoining of double strand of DNA, shown to be amplified in a subset of breast tumors with ERBB2 amplification[49].



Median Rank	p-Value	Gene
1.5	7.55E-9	ERBB2
2.5	3.17E-11	C17orf37
4.0	9.70E-37	PGAP3
4.5	1.46E-7	STARD3
5.0	7.23E-10	GRB7
15.5	6.68E-8	ORMDL3
34.0	3.63E-6	KMO
48.5	0.001	PSMD3
50.0	7.12E-5	CDK12
60.5	2.93E-4	MED1
65.5	4.33E-5	TMEM45B
80.5	3.72E-5	ABCC11
85.0	5.99E-4	GSDMB
101.5	3.63E-4	SLC22A23
109.5	0.004	PNMT
110.5	1.49E-4	TMEM86A
111.5	1.81E-4	FBXL20
114.0	3.02E-5	FHOD1
124.0	3.54E-5	CATSPERB
128.5	5.20E-5	GPCPD1

Figure 7. Top 20 genes from all chromosomes ranked according to their mRNA expression across 10 ErbB2+ breast cancer datasets using Oncomine[1]. Except for KMO, TMEM45B, SLC22A23, TMEM86A, FHOD1, CATSPERB, and GPCPD1, all the other 13 genes are from chromosome 17

The datasets are: 1) Breast Carcinoma - ERBB2 Positive, Bild Breast, Nature, 2006; 2) Ductal Breast Carcinoma - ERBB2 Positive, Bittner Breast, Not Published, 2005; 3) Ductal Breast Carcinoma Epithelia - ERBB2 Positive, Boersma Breast, Int J Cancer, 2008; 4) Ductal Breast Carcinoma - ERBB2 Positive, Bonnefoi Breast, Lancet Oncol, 2007; 5) Breast Carcinoma -ERBB2 Positive, Chin Breast, Cancer Cell, 2006; 6) Invasive Breast Carcinoma - ERBB2 Positive, Gluck Breast, Breast Cancer Res Treat, 2011; 7) Breast Carcinoma - ERBB2 Positive, Kao Breast, BMC Cancer, 2011; 8) Ductal Breast Carcinoma - ERBB2 Positive, Lu Breast, Breast Cancer Res Treat, 2008; 9) Breast Carcinoma - ERBB2 Positive, Minn Breast 2, Nature, 2005; 10) Invasive Ductal Breast Carcinoma - ERBB2 Positive, TCGA Breast, No Associated Paper, 2011.

[1] Oncomine™ (Compendia Bioscience, Ann Arbor, MI) was used for analysis and visualization.

Table 1

Cancer gene list for Chromosome 17

		Sanger List[a]	GeneCards[a]	Waldman[a]	Cancerindex[a]	Adjacent Cancer Related Genes[b]	Gene Density’
17p13.3	**ABR**			√		2+	43
	YWHAE	√				2+	51
	CRK		√	√	√	2+	51
	HIC1		√		√	2+	51
17p13.2	**USP6**	√	√		√	2 +	32
17p13.1	**ALOX12**			√		1+	47
	TP53[d]	√	√	√	√	2+	99
	PER1	√				2+	67
	RCVRN		√		√	1+	22
17p11.2	**MAP2K4**	√	√		√	1+	21
	FLCN	√				2+	66
17q11.2	**TRAF4**			√		1+	51
	NF1[d]	√	√	√	√	2+	36
17q12	**TAF15**	√			√	2+	77
	LASP1	√	√		√	2+	57
	ERBB2[d]	√	√	√	√	3 +	57
17q21.1	**THRA**		√	√	√	3+	76
17q21.2	**RARA**[d]	√	√	√	√	3+	76
	ToP2A		√	√	√	3+	76
	CCR7		√		√	3+	76
17q21.31	**BRCA1**[d]	√	√	√	√	3+	58
	NBR1		√	√	√	3+	58
	ETV4[d]	√	√	√	√	3+	58
	WNT3			√		2+	55
17q21.33	**PHB**			√		2+	60

		Sanger List[a]	GeneCards[a]	Waldman[a]	Cancerindex[a]	Adjacent Cancer Related Genes[b]	Gene Density’
	NGFR			√		2+	60
	ABCC3		√		√	2+	86
	NME1		√		√	2+	23
	NME2		√		√	2+	23
17q23.1	**CLTC**	√	√			2+	33
17q23.2	**BRIP1**	√				2+	20
17q23.3	**CD79B**	√	√		√	2+	54
	DDX5	√				2+	54
17q24.1	**AXIN2**		√		√	2+	12
17q24.2	**PRKAR1A**	√	√		√	1+	42
17q25.1	**SSTR2**		√	√	√	1+	20
	GRB2			√		1+	69
	ITGB4			√		1+	69
17q25.2	**SEPT9**	√				2+	28
17q25.3	**TK1**			√		3+	54
	BIRC5				√	3+	54
	TIMP2			√	√	3+	54
	LGALS3BP			√		3+	54
	CANT1	√				3+	54
	MAFG		√	√	√	1+	88

[a] The websites which list oncogene information are as follows: http://www.sanger.ac.uk/genetics/CGP/Census/ http://www.genecards.org/ http://waldman.ucsf.edu/GENES/completechroms.html http://www.cancerindex.org/geneweb/genes_d.htm

[b] Information derived from GeneCards. As a way to assess degree of cancer associations we have calculated a score as follows: oncogene designation + 5 points, direct literature association with cancer +3 points, present in cancer datasets +1 point. The score scale was set as follows: >=25: 3+, scores 10~25: 2+, scores <10: 1+.

[c] The gene density was derived from GeneCards. The solid bars represent the regions with a gene density above the average on chromosome 17, that is 30.3.

[d] Recognized as driver oncogene

Table 2

The 46 genes identified with more than one transcript from the six ErbB2+ cancer cell lines: two colorectal (LIM2405, LIM1899), two gastric (KATOIII, SNU16) and two breast (SUM149, SUM190).

AANAT	CARD14	GPS1	RECQL5
ABR	CASKIN2	GRB2	RNF213
ACBD4	CBX2	HIC1	Septin9
AFMID	CDK12	LLGL2	SKA2
ALOXE3	CEP112	MAPK7	SMG6
ANAPC11	CPD	MSI2	SPHK1
B4GALNT2	CTNS	MXRA7	TADA2A
BAIAP2	DLX4	NME2	THRA
C17orf57	EFCAB3	NXN	TTYH2
C17orf81	EIF5A	PDK2	UBTF
CACNB1	GAS7	PEMT	
CAMKK1	GOSR2	PHF12	

Table 3

Top over-expressed chromosome 17 genes in three human breast cancer subtypes[a,b].

ErbB2 Positive				Estrogen Receptor Positive				ErbB2/ER/PR Negative			
Gene Symbol	Band	Gene Rank	P-Value	Gene Symbol	Band	Gene Rank	P-Value	Gene Symbol	Band	Gene Rank	P-Value
ERBB2	17q12	1.5	7.55E-09	MAPT	17q21.31	32	1.00E-12	KRT16	17q21.2	147.5	5.74E-07
C17orf37	17q12	3	2.66E-11	LOC440459	17q24.1	50	9.18E-10	KIF18B	17q21.31	273	2.08E-07
PGAP3	17q12	4	5.13E-11	WNK4	17q21.31	126	1.13E-12	KRT23	17q21.2	490	1.08E-05
STARD3	17q12	5	1.12E-35	RUNDC1	17q21.31	185	1.42E-11	KRT17	17q21.2	573.5	3.37E-06
GRB7	17q12	5	7.23E-10	SLC16A6	17q24.2	186	1.35E-05	RHBDF2	17q25.1	662	2.29E-09
ORMDL3	17q12	21	1.15E-11	ARSG	17q24.2	197	3.44E-09	TRIM47	17q25.1	798	2.12E-04
PSMD3	17q21.1	48.5	1.00E-03	CHAD	17q21.33	206	5.01E-24	BIRC5	17q25.3	811	1.79E-04
CDK12	17q12	56	1.33E-04	RARA	17q21.2	258.5	3.67E-07	SPHK1	17q25.1	880	5.40E-05
MED1	17q12	60.5	2.93E-04	HEXIM1	17q21.31	267	1.76E-07	CCL18	17q12	903.5	4.65E-04
GSDMB	17q12	85	5.99E-04	IGFBP4	17q21.2	267.5	9.02E-08	1CAM2	17q23.3	944	1.00E-03
PNMT	17q12	137	7.00E-03	SPATA20	17q21.33	346	1.95E-10	SEC14L1	17q25.2	1024	6.39E-04
TCAP	17q12	148	2.24E-07	LRRC46	17q21.32	377	1.42E-07	PRKCA	17q24.2	1121.5	3.00E-03
FBXL20	17q12	164	3.61E-04	STH	17q21.31	411.5	1.04E-06	KRT14	17q21.2	1273.5	1.00E-03
CYB561	17q23.3	187	2.00E-03	BECN1	17q21.31	436.5	3.45E-07	CCL7	17q12	1288	8.90E-04
MED24	17q21.1	202.5	1.00E-03	NPEPPS	17q21.32	468	3.90E-06	CCL2	17q12	1308	3.71E-04
SLC39A11	17q25.1	254	1.55E-04	G6PC3	17q21.31	471	7.10E-17	CCL5	17q12	1322	2.00E-03
TLCD1	17q11.2	383	7.18E-07	SLC9A3R1	17q25.1	472	2.00E-02	FAM100B	17q25.1	1382	3.00E-03
SRCIN1	17q12	389	4.24E-04	CACNG4	17q24.2	504	2.73E-09	STAC2	17q12	1419.5	5.25E-04
TANC2	17q23.2	465	2.00E-03	C17orf58	17q24.2	514	1.80E-08	MAFG	17q25.3	1436	7.27E-09

[a] The rank for a gene is the median rank for that gene across each of the analyses. The p-value for a gene is its p-value for the median-ranked analysis.

[b] The genes are listed in chromosome regions to highlight the association of gene regions with breast cancer phenotypes.