Supplementary materials for MetaDCN: meta-analysis framework for differential co-expression network detection with an application in breast cancer

Li Zhu [1†], Ying Ding [2†], Cho-Yi Chen [1,3,†], Lin Wang [1], Zhiguang Huo [1], SungHwan Kim [1], Christos Sotiriou [4], Steffi Oesterreich [5], and George C. Tseng [1,2,*]

[1]Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, PA, 15261, U.S.; [2]Department of Computational and Systems Biology, School of Medicine, University of Pittsburgh, Pittsburgh, PA, 15213, U.S.; [3]Genome and Systems Biology Degree Program, National Taiwan University, Taipei, 10617, Taiwan; [4]Breast Cancer Translational Research Laboratory, J. C. Heuson, Institut Jules Bordet, University Libre de Bruxelles, Brussels 1000, Belgium; [5]Magee-Women's Research Institute, Pittsburgh, PA, 15213, U.S.
† Those authors contributed equally
* To whom correspondence should be addressed. Email: ctseng@pitt.edu

1 MetaDCNExplorer algorithm

MetaDCNExplorer is a Cytoscape application (App) for visualization of differential co-expression networks (DCNs). MetaDNCExplorer utilizes the power of Cytoscape Java API to visualize complex networks. The graphical user interface (GUI) allows users to load input network files and any node/edge attribute tables associated with the networks. Users can manage the imported networks and all aesthetic elements via control panel. MetaDCNExplorer was designed to generate visualization for differential co-expression networks in which nodes represent genes and edges represent co-expression relationships. Each edge should be associated with following two attributes: 1) the directional effect size (e.g., Z-score) and 2) the statistical significance (P-value) of differential co-expression. A gene can belong to one or many modules in a network. Node attributes should specify the module membership of the gene. All above-mentioned attributes, along with the modular network they attached to, are necessary for MetaDCNExplorer, and can be automatically generated from the analysis pipeline of MetaDCN R package. MetaNetworkExplorer was developed in Java programming language and built on OSGi (Open Service Gateway Initiative) Java framework. The implement was based on Cytoscape archetype cyaction-app version 3.0.0 and was built as Bundle App that can be dynamically loaded by Cytoscape main program. By default the prefuse force-directed layout was used to visualize the modular structure hidden in the input network. This layout is based on the force simulation algorithm implemented as part of the prefuse toolkit (Heer *et al.*, 2005), integrated in the Cytoscape main program. The algorithm positions nodes based on a physics simulation of interacting forces that consist of node repelling force, edge spring force, and air drag forces. The absolute effect size of differential co-expression (i.e., Z-score) reflects the spring length in the simulation. Inter-module repelling factor and intra-module attracting factor is provided for tuning. User can also select either linear force or exponential force. The estimated running time of this layout algorithm on a network with N nodes and E edges will be the grater of O(N log N) and O(E).

2 Data description and preprocessing

Eight breast cancer datasets (five training sets and three testing sets) were used for comparing ER+ and ER- patients, including six GEO datasets, The Cancer Genome Atlas (TCGA) breast cancer dataset, and Molecular Taxonomy of Breast Cancer International Consortium dataset (METABRIC) (see Table S1). The TCGA breast cancer dataset was downloaded from the Cancer Genome Atlas (TCGA) website (http://tcga-data.nci.nih.gov/tcga) in October 2012. Level 3 RNA-Seq data were extracted from the Illumina HiSeq 2000 platform. We selected the TCGA breast cancer dataset that contained expression data of n=406 tumor samples. The METABRIC gene expression and clinical data were retrieved from Synapse (https://www.synapse.org/#!Synapse:syn2133309) where we obtained 1981 samples (Curtis *et al.*, 2012). In all studies, microarrays were scanned and summarized by manufacturers' defaults. For the six studies from GEO, data from Affymetrix arrays were processed by robust multi-array (RMA) method and data from Illumina arrays by manufacturer's BeadArray software for probe analysis. Oligonucleotide probes (or probesets) were matched to gene symbols using hgu133plus2.db and illuminaHumanv4.db Bioconductor packages. If multiple probes matched to the same gene, the probe with the largest inter-quantile range (IQR) was used. After matching all the genes across the eight studies, we further filtered away genes with average standard deviation smaller than 0.2 across all studies, which left 10,636 genes for the following analysis.

Four breast cancer datasets (2 training sets and 2 testing sets) were used for comparing invasive lobular carcinoma (ILC) and invasive ductal carcinoma (IDC) (see Table S2). We included ILC and IDC of Lumina A subtypes from METABRIC and TCGA datasets to gain better homogeneity in patients. TCGA transcript per million (TPM) data were achieved from GSE62944 (Rahman *et al.*, 2015). PAM50 (Parker *et al.*, 2009) subtypes of TCGA patients were called by applying genefu R package (Haibe-Kains *et al.*, 2012), using an ER balanced subsamples for median centering (Curtis *et al.*, 2012). We also included ILCs in a dataset from Sotiriou Lab (Metzger-Filho *et al.*, 2013) and a dataset from Rational Therapy for Breast Cancer (RATHER) consortium (Michaut *et al.*, 2016) excluding overlapping patients in METABRIC, for validation. The pre-processing step is similar to the previous section. After matching genes and filtering out all the genes with average gene expression or average standard deviation smaller than 50% across two studies, 4552 genes left for following analysis.

All these studies were approved by the University of Pittsburgh Institutional Review Board (IRB PRO16020311).

3 Supplementary tables and figures

Table S1: Description of breast cancer datasets for comparing ER+ vs. ER-

	Data sets	Sample size (ER+ vs. ER-)	Platform
Training	TCGA (S1)	406(319 vs. 87)	RNA-Seq
	GSE7390 (S2)	198(134 vs. 64)	Affymetrix HG-U133A
	GSE2034 (S3)	286(209 vs. 77)	Affymetrix HG-U133A
	METABRIC (S4)	1981(1512 vs. 469)	Illumina
	GSE4922 (S5)	245(211 vs. 34)	Affymetrix HG-U133A
Testing	GSE23720 (S6)	197(131 vs. 66)	Affymetrix HG-U133 Plus 2.0
	GSE58215 (S7)	270(218 vs. 52)	Agilent-028004
	GSE22220 (S8)	216(134 vs. 82)	Illumina humanRef8

Table S2: Description of breast cancer datasets for comparing ILC vs. IDC

	Data sets	Sample size (ILC vs. IDC)	Platform
Training	TCGA	470 (159 vs. 311)	RNA-Seq
	METABRIC	598 (65 vs. 533)	Illumina
Testing	Sotiriou	147 ILCs	Affymetrix HG-U133 Plus 2.0
	RATHER	111 ILCs	Agilent custom-designed platform

Table S3: Pathway-centric supermodules with at least 3 pathway overlapping genes (with 10 repeats with different initial modules). Module starts with H indicates it is more densely connected in ILC network; module starts with L indicates it is more densely connected in IDC network.

Pathway name (ILC vs. IDC)	Pathway size	Module size	# pathway genes	q-value	p-value	Module
GO_PROTEASE_INHIBITOR_ACTIVITY	41	27	3	3.0e-03	6.1e-05	L2,L4,L8
GO_PROTEINACEOUS_EXTRACELLULAR_MATRIX	98	15	3	3.0e-03	8.5e-04	L5,L7
GO_EXTRACELLULAR_MATRIX	100	15	3	3.0e-03	8.5e-04	L5,L7
GO_REGULATION_OF_CELL_CYCLE	182	22	3	1.8e-02	6.8e-03	L4,L8
KEGG_FOCAL_ADHESION	201	22	3	2.7e-02	1.3e-02	L7,L8

Table S4: Pathway-centric supermodules with at least 3 pathway overlapping genes (with 10 repeats with different initial modules). Module starts with H indicates it is more densely connected in ER+ network; module starts with L indicates it is more densely connected in ER- network.

Pathway name (ER+ vs. ER-)	Pathway size	Module size	# pathway genes	q-value	p-value	Module
REACTOME_COMPLEMENT_CASCADE	32	25	4	2.1e-05	1.9e-07	H7,H8
GO_IMMUNE_RESPONSE	235	28	7	5.6e-05	1.3e-06	H9,L1,L2
REACTOME_REGULATION_OF_COMPLEMENT_CASCADE	14	25	3	5.6e-05	2.5e-06	H7,H8
GO_ORGAN_MORPHOGENESIS	144	35	6	5.6e-05	2.8e-06	H3,H5,L9
BIOCARTA_TCYTOTOXIC_PATHWAY	14	23	3	5.6e-05	2.8e-06	H3,L5
GO_PROTEINACEOUS_EXTRACELLULAR_MATRIX	98	28	5	5.6e-05	3.9e-06	H1,H5
GO_EXTRACELLULAR_MATRIX	100	28	5	5.6e-05	3.9e-06	H1,H5
BIOCARTA_THELPER_PATHWAY	14	23	3	5.6e-05	4.1e-06	H3,L5
KEGG_COMPLEMENT_AND_COAGULATION_CASCADES	69	25	4	1.1e-04	8.8e-06	H7,H8
REACTOME_CELL_SURFACE_INTERACTIONS_AT_THE_VASCULAR_WALL	91	31	5	1.1e-04	9.8e-06	H7,L1,L5
KEGG_SYSTEMIC_LUPUS_ERYTHEMATOSUS	140	21	4	1.5e-04	1.4e-05	H8,H9
KEGG_PROTEASOME	48	17	3	1.5e-04	1.7e-05	H13,L1
GO_EXTRACELLULAR_MATRIX_STRUCTURAL_CONSTITUENT	27	23	3	1.5e-04	1.9e-05	H2,H6
REACTOME_CDK_MEDIATED_PHOSPHORYLATION_AND_REMOVAL_OF_CDC6	48	17	3	1.5e-04	2.3e-05	H13,L1
REACTOME_CDT1_ASSOCIATION_WITH_THE_CDC6_ORC_ORIGIN_COMPLEX	56	17	3	1.5e-04	2.3e-05	H13,L1
REACTOME_CROSS_PRESENTATION_OF_SOLUBLE_EXOGENOUS_ANTIGENS_ENDOSOMES	48	17	3	1.5e-04	2.6e-05	H13,L1
REACTOME_AUTODEGRADATION_OF_THE_E3_UBIQUITIN_LIGASE_COP1	51	17	3	1.5e-04	2.6e-05	H13,L1
GO_REGULATION_OF_NEUROGENESIS	14	38	3	1.5e-04	2.6e-05	H3,H5,H7
REACTOME_REGULATION_OF_ORNITHINE_DECARBOXYLASE_ODC	49	17	3	1.5e-04	2.9e-05	H13,L1
REACTOME_P53_INDEPENDENT_G1_S_DNA_DAMAGE_CHECKPOINT	51	17	3	1.5e-04	2.9e-05	H13,L1
REACTOME_SCF_BETA_TRCP_MEDIATED_DEGRADATION_OF_EMI1	51	17	3	1.5e-04	2.9e-05	H13,L1
REACTOME_DESTABILIZATION_OF_MRNA_BY_AUF1_HNRNP_D0	53	17	3	1.5e-04	2.9e-05	H13,L1
GO_HUMORAL_IMMUNE_RESPONSE	32	19	3	1.6e-04	3.2e-05	H9,L1
REACTOME_VIF_MEDIATED_DEGRADATION_OF_APOBEC3G	52	17	3	1.7e-04	3.7e-05	H13,L1
REACTOME_SCFSKP2_MEDIATED_DEGRADATION_OF_P27_P21	56	17	3	2.1e-04	4.6e-05	H13,L1
GO_GLYCOSAMINOGLYCAN_BINDING	34	21	3	2.1e-04	5.0e-05	H3,L12,L14
GO_POLYSACCHARIDE_BINDING	36	21	3	2.3e-04	5.7e-05	H3,L12,L14
REACTOME_COSTIMULATION_BY_THE_CD28_FAMILY	63	14	3	2.3e-04	5.7e-05	L2,L5
REACTOME_IL_3_5_AND_GM_CSF_SIGNALING	43	21	3	3.3e-04	9.0e-05	H9,L2
GO_LIPID_RAFT	29	34	3	3.8e-04	1.1e-04	H7,H9,L1
REACTOME_COLLAGEN_FORMATION	58	24	3	7.2e-04	2.4e-04	H1,H3
GO_GENERATION_OF_NEURONS	83	25	3	7.2e-04	2.5e-04	H3,H5
GO_NEUROGENESIS	93	25	3	7.3e-04	2.7e-04	H3,H5
KEGG_T_CELL_RECEPTOR_SIGNALING_PATHWAY	108	14	3	8.4e-04	3.3e-04	L2,L5
KEGG_NATURAL_KILLER_CELL_MEDIATED_CYTOTOXICITY	137	14	3	1.0e-03	4.2e-04	L2,L5
GO_STRUCTURAL_MOLECULE_ACTIVITY	244	22	4	1.5e-03	7.1e-04	H5,H8
GO_RESPONSE_TO_WOUNDING	190	27	4	1.6e-03	8.1e-04	H5,H7
REACTOME_EXTRACELLULAR_MATRIX_ORGANIZATION	87	24	3	1.6e-03	8.6e-04	H1,H3
GO_NEGATIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	197	27	3	1.3e-02	9.5e-03	H5,H7
GO_CATION_BINDING	213	24	3	1.3e-02	9.8e-03	H4,H5

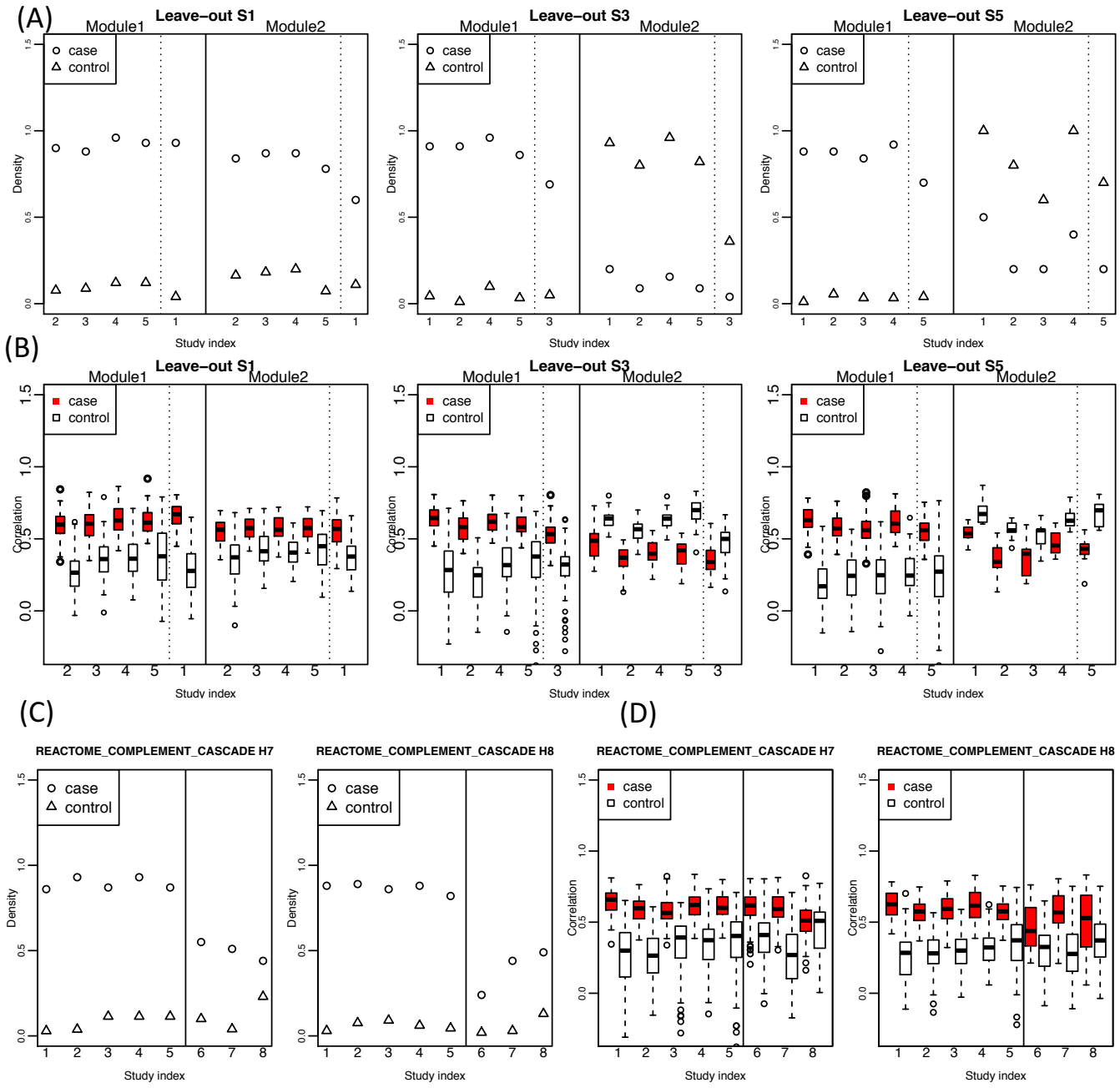

Figure S1: (A) Densities and (B) correlations of the basic modules assembled into complement cascade pathway supermodules in leave-one-out cross-validation. Leaving out study 2 and 4 do not give supermodules significant enriched in complement cascade pathway. Solid lines separate modules, and dashed lines separate training set and testing set. (C) Module density and (D) correlations of genes in the basic modules enriched in complement cascade supermodule in independent validation studies. Solid lines separate training sets and testing sets.

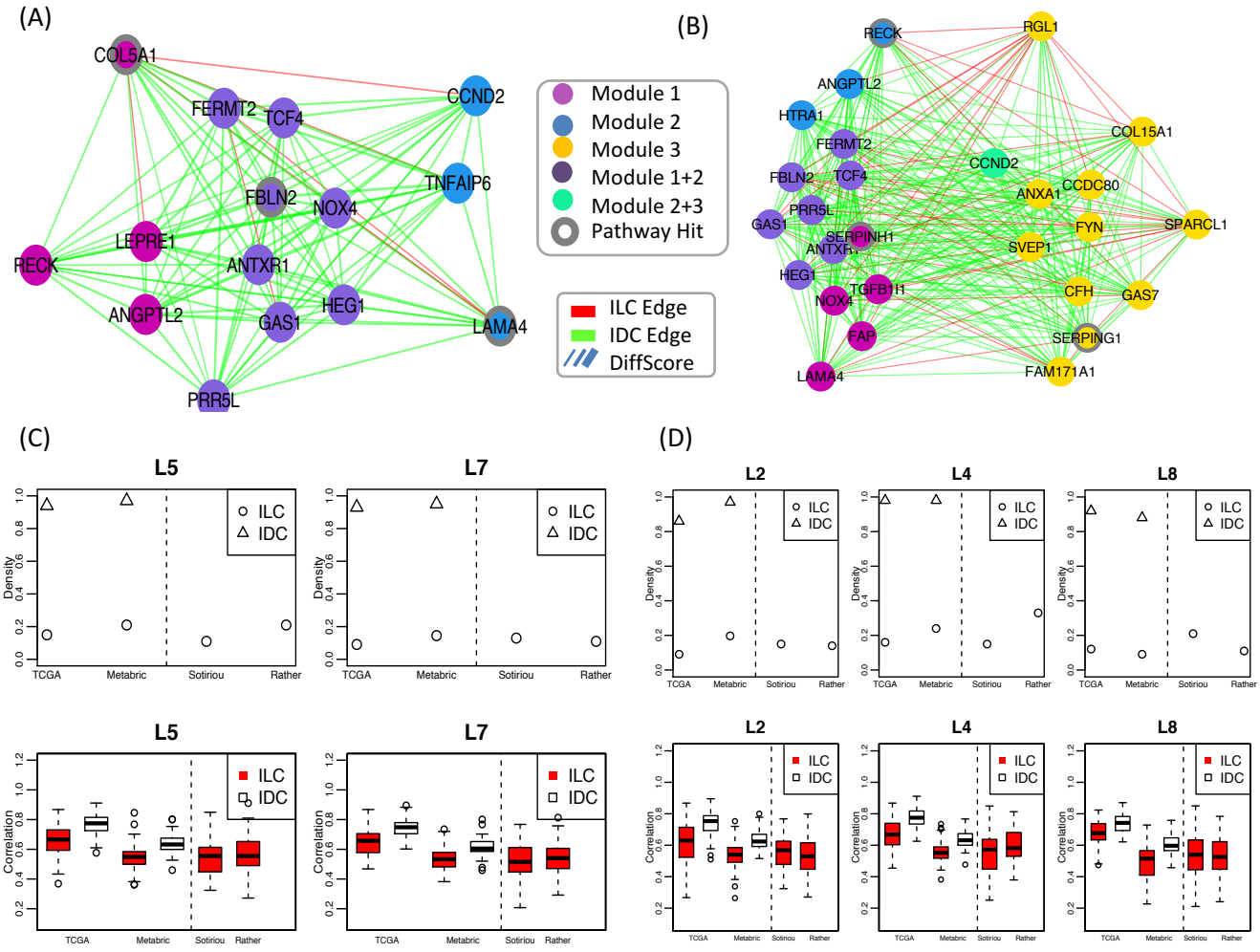

Figure S2. (A) Visualization of proteinaceous extracellular matrix pathway supermodule. (B) Visualization of protease inhibitor activity pathway supermodules. The edge color represents the direction of differential gene co-expression, in which the positive value implies ILC-favored co-expression and the negative value implies IDC-favored co-expression. Node color represents its origin of sub-modules, and the genes annotated in the immune response pathway are highlighted with dark circles. Edge width represents edge weight (Z score of differential co-expression). (C) Module densities and (D) gene-gene pairwise correlations of the basic modules enriched in those two pathways in TCGA, METABRIC, Sotiriou and RATHER. Dashed lines separate training and testing sets.

References

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Langerød, A., Green, A., Provenzano, E., Wishart, G., Pinder, S., Watson, P., Markowetz, F., Murphy, L., Ellis, I., Purushotham, A., Børresen-Dale, A.-L., Brenton, J. D., Tavaré, S., Caldas, C., and Aparicio, S. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403), 346–52.

Haibe-Kains, B., Desmedt, C., Loi, S., Culhane, A. C., Bontempi, G., Quackenbush, J., and Sotiriou, C. (2012). A three-gene model to robustly identify breast cancer molecular subtypes. *Journal of the National Cancer Institute*, **104**(4), 311–325.

Heer, J., Card, S. K., and Landay, J. A. (2005). Prefuse: A toolkit for interactive information visualization. In *In ACM Human Factors in Computing Systems (CHI)*, pages 421–430.

Metzger-Filho, O., Michiels, S., Bertucci, F., Catteau, A., Salgado, R., Galant, C., Fumagalli, D., Singhal, S. K., Desmedt, C., Ignatiadis, M., Haussy, S., Finetti, P., Birnbaum, D., Saini, K. S., Berlière, M., Veys, I., de Azambuja, E., Bozovic, I., Peyro-Saint-Paul, H., Larsimont, D., Piccart, M., and Sotiriou, C. (2013). Genomic grade adds prognostic value in invasive lobular carcinoma. *Annals of Oncology*, **24**(2), 377–384.

Michaut, M., Chin, S.-F., Majewski, I., Severson, T. M., Bismeijer, T., de Koning, L., Peeters, J. K., Schouten, P. C., Rueda, O. M., Bosma, A. J., Tarrant, F., Fan, Y., He, B., Xue, Z., Mittempergher, L., Kluin, R. J., Heijmans, J., Snel, M., Pereira, B., Schlicker, A., Provenzano, E., Ali, H. R., Gaber, A., O'Hurley, G., Lehn, S., Muris, J. J., Wesseling, J., Kay, E., Sammut, S. J., Bardwell, H. A., Barbet, A. S., Bard, F., Lecerf, C., O'Connor, D. P., Vis, D. J., Benes, C. H., McDermott, U., Garnett, M. J., Simon, I. M., Jirström, K., Dubois, T., Linn, S. C., Gallagher, W. M., Wessels, L. F., Caldas, C., and Bernards, R. (2016). Integration of genomic, transcriptomic and proteomic data identifies two biologically distinct subtypes of invasive lobular breast cancer. *Scientific Reports*, **6**(November 2015), 18517.

Parker, J. S., Mullins, M., Cheung, M. C. U., Leung, S., Voduc, D., Vickery, T., Davies, S., Fauron, C., He, X., Hu, Z., Quackenhush, J. F., Stijleman, I. J., Palazzo, J., Matron, J. S., Nobel, A. B., Mardis, E., Nielsen, T. O., Ellis, M. J., Perou, C. M., and Bernard, P. S. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology*, **27**(8), 1160–1167.

Rahman, M., Jackson, L. K., Johnson, W. E., Li, D. Y., Bild, A. H., and Piccolo, S. R. (2015). Alternative preprocessing of RNA-Sequencing data in the Cancer Genome Atlas leads to improved analysis results. *Bioinformatics*, **31**(22), 3666–3672.