www.sciencetranslationalmedicine.org/cgi/content/full/5/181/181re1/DC1

Supplementary Materials for

Systematic Analysis of Challenge-Driven Improvements in Molecular Prognostic Models for Breast Cancer

Adam A. Margolin,* Erhan Bilal, Erich Huang, Thea C. Norman, Lars Ottestad, Brigham H. Mecham, Ben Sauerwine, Michael R. Kellen, Lara M. Mangravite, Matthew D. Furia, Hans Kristian Moen Vollan, Oscar M. Rueda, Justin Guinney, Nicole A. Deflaux, Bruce Hoff, Xavier Schildwachter, Hege G. Russnes, Daehoon Park, Veronica O. Vang, Tyler Pirtle, Lamia Youseff, Craig Citro, Christina Curtis, Vessela N. Kristensen, Joseph Hellerstein, Stephen H. Friend,* Gustavo Stolovitzky, Samuel Aparicio, Carlos Caldas, Anne-Lise Børresen-Dale

*Corresponding author. E-mail: margolin@sagebase.org (A.A.M.); friend@sagebase.org (S.H.F.)

Published 17 April 2013, *Sci. Transl. Med.* **5**, 181re1 (2013)
DOI: 10.1126/scitranslmed.3006112

The PDF file includes:

SUPPLEMENTARY MATERIALS

Materials and Methods

OsloVal data generation

Patient material: The OsloVal cohort consists of fresh frozen primary tumors from 184 breast cancer patients collected from 1981 to 99 (148 from 1981-89 and 36 from 1994-99) at the Norwegian Radium Hospital. Tumor material was collected for hormone receptor analysis by DCC (dextran coated charcoal) and excess tumor material was stored in a biobank (at -80^{o}C). Frozen tissue samples were first cut in three parts, and tissue sections made from each (6 µm) for later evaluations using IHC. Tissue-Tek was removed from the specimens and all tissue were sliced, mixed and homogenized, and tissue aliquots stored at -80^{o}C. Two tissue fractions of approximately 15 mg each were used for DNA and RNA isolation and analyses.

mRNA extraction: Extractions were performed on the QIA symphony SP robot from Qiagen. A total of 400 µl RLT buffer was added to the samples while on dry ice, followed by homogenization (Tissuelyzer). The QIA symphony RNA Kit cat# 931636 from Qiagen was used. Extracted RNA was quantified using Nanodrop 1000 and the RNA integrity determined using the Agilent Bioanalyzer.

DNA extraction: Extractions were performed on the QIAsymphony SP robot from Qiagen. A total of 20µl Proteinase K and 180µl ATL buffer was added to the tissue aliquot and lysed for three hours at 56 degrees on a Thermomixer with 900 rpm, followed by addition of 4 µl RNase A and incubation at RT for 2 min. The extraction kit used was the QIAsymphony DNA mini Kit cat#931236 from Qiagen,

CNA analyses: A total of 500 ng Genomic DNA was used for the Affymetrix Genome-Wide Human SNP 6.0 Assay, and processed according to the manufacturer's protocol using the Affymetrix GW Human SNP Nsp/Sty 6.0 assay kit (prod. # 901015). Fragmented and end labeled PCR product were hybridized to the Affymetrix Human SNP 6.0 array (prod. # 901150) followed by incubation for 16 h at 50^0C at constant rotation (60 rpm). The washing and staining procedure was performed in the Affymetrix Fluidics Station 450. The arrays were scanned at 560 nm using a confocal laser-scanning microscope (Affymetrix Scanner 3000 7G). The SNP calls and Copy number analysis were done using the Affymetrix Genotyping Console Software ver 4.0.

Expression analysis: A total of 100 ng total RNA was used for cRNA preparation using the Ambion Illumina Total Prep kit (prod. # AMIL1791). The samples were hybridized to Illumina HT-12 v4 BeadChips, incubated 16h at 580C and subsequently scanned using the Illumina BeadStation.

Clinical and pathology variables: Routine tissue sections were retrieved for all patient samples (hematoxylin and eosin stained). Histological type was assessed according to WHO classification *(45)* and then histological grade were determined *(46)*. ER and PR status were scored using mRNA expression data. HER2 amplification status was scored from SNP6 arrays. Survival time was collected from the national registries

Data availability

We provide the METABRIC and OsloVal data described in the paper via two mechanisms. Each mechanism is under different Terms of Use described in the links, be cognizant that these are patient data and therefore fall under data governance rules:

(A) For a 6-month post-Challenge "Validation Phase", we provide registered participants access to the data to re-evaluate their models

https://synapse.prod.sagebase.org/#!Synapse:syn1710250

(B) For those interested in accessing these data for independent research, please use these links for further information on how to obtain access to these protected health data

METABRIC: https://synapse.prod.sagebase.org/#!Search:syn1688369

OsloVal: https://synapse.prod.sagebase.org/#!Search:syn1688370

Table S1: Univariate and multivariate Cox regression statistics for the clinical covariates in the METABRIC data set.

	Multivariate Cox regression					Univariate Cox regression				
Covariate	coef	exp(coef)	se(coef)	z	Pr(>|z|)	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Age at diagnosis	0.03946	1.040249	0.00349	11.3074	0	0.028595	1.029008	0.002929	9.764013	0
Tumor size	0.010088	1.010139	0.001815	5.557924	2.73E-08	0.014921	1.015032	0.001511	9.872802	0
Lymph nodes	0.053953	1.055435	0.006378	8.459218	0	0.06636	1.068612	0.004943	13.42513	0
Intermediate grade	0.217714	1.243232	0.150131	1.450162	0.147013	0.294219	1.342078	0.148386	1.982796	0.04739
High grade	0.313008	1.367533	0.152654	2.050437	0.040322	0.604817	1.830917	0.144895	4.174163	2.99E-05
ER negative	0.145647	1.156787	0.115006	1.266429	0.205359	0.320903	1.378372	0.076893	4.173364	3.00E-05
PR negative	0.145485	1.1566	0.08454	1.720888	0.085271	0.369708	1.447311	0.067268	5.496077	3.88E-08
HER2 negative	-0.25087	0.778123	0.103818	-2.41646	0.015672	-0.46318	0.62928	0.093925	-4.93135	8.17E-07
CT/HT	-0.08126	0.921953	0.334094	-0.24323	0.807829	-0.22345	0.799753	0.319135	-0.70018	0.483813
CT/HT/RT	-0.77431	0.461022	0.22377	-3.4603	0.00054	-0.55967	0.571396	0.215676	-2.59497	0.00946
CT/RT	-0.35825	0.698896	0.21013	-1.70491	0.088211	-0.27657	0.758382	0.204537	-1.35217	0.176322
HT	-0.906	0.404138	0.215142	-4.21115	2.54E-05	-0.56977	0.565655	0.187623	-3.03678	0.002391
HT/RT	-1.0773	0.340514	0.210783	-5.11095	3.21E-07	-0.76561	0.465052	0.18511	-4.13596	3.53E-05
RT	-1.24311	0.288487	0.226251	-5.49438	3.92E-08	-1.32218	0.266553	0.207634	-6.36785	1.92E-10
No treatment	-0.89846	0.407194	0.217971	-4.12194	3.76E-05	-0.86439	0.421307	0.193406	-4.46932	7.85E-06

Treatment:
CT - Chemo therapy
RT - Radio therapy
HT - Hormonal therapy

Table S2: Univariate and multivariate Cox regression statistics for the clinical covariates in the OsloVal data set.

	Multivariate Cox regression					Univariate Cox regression				
Covariate	coef	exp(coef)	se(coef)	z	Pr(>|z|)	coef	exp(coef)	se(coef)	z	Pr(>|z|)
Age at diagnosis	0.038351	1.039096	0.015439	2.483999	0.012992	0.034676	1.035285	0.007749	4.475184	7.63E-06
Tumor size	0.086823	1.090704	0.119527	0.726388	0.467601	0.166058	1.180641	0.060727	2.73449	0.006248
Lymph nodes	0.145669	1.156814	0.034743	4.192749	2.76E-05	0.132847	1.142075	0.017272	7.69146	1.45E-14
Intermediate grade	0.296181	1.344713	0.838911	0.353054	0.724048	0.775138	2.170891	0.470543	1.647326	0.099491
High grade	0.568608	1.765807	0.848445	0.670177	0.502745	0.846257	2.330906	0.474803	1.782331	0.074695
ER negative	0.947331	2.578818	0.409886	2.311205	0.020822	0.298188	1.347415	0.177244	1.682354	0.0925
PR negative	-0.22781	0.796275	0.406875	-0.5599	0.575546	0.311458	1.365415	0.219119	1.421411	0.155197
HER2 negative	0.221261	1.24765	0.46046	0.480522	0.630856	-0.19023	0.826768	0.254062	-0.74876	0.454002
CT/HT	0.677686	1.969316	0.43508	1.557614	0.119325	0.902983	2.466951	0.294647	3.064628	0.002179
CT/HT/RT	0.346296	1.413821	0.566452	0.611342	0.540973	1.040813	2.831519	0.317192	3.281334	0.001033
CT/RT	-0.43961	0.644287	0.812231	-0.54124	0.588343	-0.09935	0.905427	0.609823	-0.16291	0.870586
HT	0.240398	1.271756	0.616252	0.390097	0.696464	1.240686	3.457985	0.370471	3.348945	0.000811
HT/RT	-16.0705	1.05E-07	4503.05	-0.00357	0.997153	-0.49426	0.610025	1.019176	-0.48496	0.627707
RT	0.168161	1.183127	1.12635	0.149297	0.881319	0.13017	1.139022	0.73491	0.177123	0.859412

Treatment:
CT - Chemo therapy
RT - Radio therapy
HT - Hormonal therapy

Table S3: The Breast Cancer Challenge Consortium: Challenge participants who submitted a model to phase 3 of the BCC.

First Name	Last Name	Affiliation	Email Address
Miika	Ahdesmäki	Almac Diagnostics, Almac Group, Craigavon, BT63 5QD, United Kingdom	miika.ahdesmaki@almacgroup.com
Robert	Atlas	Department of Computer Sciences, University of Wisconsin, Madison, WI 53717, USA	ratlas@cs.wisc.edu
Nikolay	Balov	Department of Biostatistics and Computational Biology University of Rochester Medical Center, Rochester, NY 14642, USA	nikibalov@gmail.com
Bonnie	Berger	Departments of Mathematics and Electrical Engineering & Computer Science Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02139, USA	bab@csail.mit.edu
Archit	Bhise	Massachusetts Institute of Technology, Cambridge, MA 02139-4307, USA	archit@MIT.EDU
Eric	Bonnet	(1) Institut Curie, 26 rue d'Ulm, Paris, F-75248 France (2) INSERM, U900, Paris, F-75248 France (3) Centre for Computational Biology - CBIO, Mines ParisTech, 35 rue Saint-Honoré, Fontainebleau, F-77300 France	eric.bonnet@curie.fr
Aaron	Boudreau	Department of Laboratory Medicine, University of California, San Francisco, San Francisco, CA 94143, USA	aaron.boudreau@ucsf.edu
Chunhui	Cai	Department of Biomedical Informatics, University of Pittsburgh, PA 15206-3701, USA	chunhuic@pitt.edu
Yifei	Chen	Department of Computer Science, University of California, Irvine, Irvine, CA 92697 USA	yifeic@uci.edu
Jie	Cheng	Quantitative Sciences, GlaxoSmithKline, Collegeville, PA 19426, USA	jcheng88@gmail.com
Sean	Cory	Goodman Cancer Research Centre, McGill University, 1160 Pine Avenue West, Montreal, Quebec, H3A 1A3, Canada	sean.cory@gmail.com
Edmund J.	Crampin	Melbourne School of Engineering, The University of Melbourne, Parkville, Victoria, Australia	e.crampin@auckland.ac.nz
Chad J.	Creighton	Division of Biostatistics, Dan L. Duncan Cancer Center, Baylor College of Medicine, Houston, Texas	creighto@bcm.edu
Benjamin	Haibe-Kains	Institut de Recherches Cliniques de Montréal, IRCM, Montreal, Quebec, Canada	bhaibeka@ircm.qc.ca
Thomas	Kelder	TNO, Microbiology and Systems Biology, Zeist, The Netherlands	thomaskelder@gmail.com
Jeff	Knisley	Institute for Quantitative Biology East Tennessee State University, Johnson City, TN 37614, USA	knisleyj.etsu@gmail.com
Vincenzo	Lagani	Bioinformatics Laboratory, Institute of Computer Science - Foundation for Research and Technology (FORTH), N. Plastira 100 Vassilika Vouton, GR-700 13 Heraklion, Crete, Greece	vlagani@ics.forth.gr
Kai Yeung	Lau	Amgen Inc, Seattle, 98119 WA, USA	kaiyeung.lau@gmail.com
Xinghua	Lu	Department of Biomedical Informatics, University of Pittsburgh, PA 15206-3701, USA	xinghua@pitt.edu
Songjian	Lu	Department of Biomedical Informatics, University of Pittsburgh, PA 15206-3701, USA	songjian@pitt.edu
Jian	Peng	Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology,, Cambridge, MA 02139, USA	jpeng@csail.mit.edu
Robert J.	Prill	IBM Almaden Research Center, San Jose, CA 95120, USA	rjprill@us.ibm.com
Markus	Ringnér	Department of Oncology, Clinical Sciences, Lund University, Lund, Sweden	markus.ringner@med.lu.se
Richard S.	Savage	Systems Biology Centre, University of Warwick, United Kingdom	r.s.savage@warwick.ac.uk
Ben	Silva	Department of Statistics, Florida State University, Tallahassee, FL 32306-4330, USA	bsilva@stat.fsu.edu
Piotr	Sobczyk	ICM, University of Warsaw, Poland	sobbombo@gmail.com
Artem	Sokolov	Baskin School of Engineering, University of California, Santa Cruz, Santa Cruz, CA 95064, USA	sokolov@soe.ucsc.edu
Emmett	Sprecher	Dept of Pathology Informatics, Yale University, New Haven, CT 06511	emmett.sprecher@yale.edu
Ioannis	Tsamardinos	(1) Bioinformatics Laboratory, Institute of Computer Science - Foundation for Research and Technology (FORTH) N. Plastira 100 Vassilika Vouton, GR-700 13 Heraklion, Crete, Greece (2) Department of Computer Science, University of Crete P.O.Box 2208, GR-714 09 Heraklion, Crete, Greece	tsamard@ics.forth.gr
Jean-Philipp	Vert	(1) Centre for Computational Biology - CBIO, Mines ParisTech, 35 rue Saint-Honoré, Fontainebleau, F-77300 France (2) Institut Curie, 26 rue d'Ulm, Paris, F-75248 France (3) INSERM, U900, Paris, F-75248 France	jean-philippe.vert@mines.org
Yi Kan	Wang	Auckland Bioengineering Institute, The University of Auckland, New Zealand	yikan.wang@auckland.ac.nz
Charles D.	Warden	Bioinformatics Core, Department of Molecular Medicine, City of Hope National Medical Center, Duarte, CA, 91010, USA	cwarden@coh.org
Xiaohui	Xie	Department of Computer Science, University of California, Irvine, Irvine, CA 92697 USA	xhx@ics.uci.edu