Clinical Value of RNA Sequencing-Based Classifiers for Prediction of the Five Conventional Breast Cancer Biomarkers: A Report From the Population-Based Multicenter Sweden Cancerome Analysis Network-Breast Initiative

Christian Brueffer, MSc, Johan Vallon-Christersson, PhD, Dorthe Grabau, MD PhD, Anna Ehinger, MD, Jari Häkkinen, PhD, Cecilia Hegardt, PhD, Janne Malina, MD, Yilun Chen, MSc, Pär-Ola Bendahl, PhD, Jonas Manjer, MD PhD, Martin Malmberg, MD PhD, Christer Larsson, PhD, Niklas Loman, MD PhD, Lisa Rydén, MD PhD, Åke Borg, PhD, and Lao H. Saal, MD PhD

SUPPLEMENTARY METHODS

SUPPLEMENTARY METHODS

Patients

The study schema is presented in Figure 1. The study was approved by the Regional Ethical Review Board of Lund at Lund University (diary numbers 2007/155, 2009/658, 2009/659, 2010/383, 2012/58, 2013/459) and the Swedish Data Inspection group (364-2010). Trained health professionals provided patient information and patients gave written informed consent. Clinical records were retrieved from the Swedish National Cancer Registry (NKBC). Diagnostic pathology slides, snap-frozen surgical tumor specimens, and formalin-fixed paraffin-embedded (FFPE) tissue blocks were retrieved for 405 patients (training cohort; Supplementary Table 1) diagnosed between 2006 and 2010 and treated at the Skåne University Hospital in Malmö and Lund. The 405 cohort was assembled for classifier training purposes and not for survival analysis, and thus an overrepresentation of HER2+ and ER– cases was selected for.

Independent validation cohort

For testing of the classifiers and for survival analyses, an independent, prospective, and population-based cohort of 3273 primary breast tumors, diagnosed between September 2010 and March 2015, was assembled from the ongoing SCAN-B study[1] (validation cohort; Supplementary Table 1 and Supplementary Figure 1).

Histopathology

For the 405 training cases, all biomarkers with the exception of Ki67 had been evaluated at time of diagnosis. The original clinical diagnostic pathology slides and scores were retrieved. For ER and PgR, the diagnostic IHC results were classified into the categories 0%, 1-10%, 11-50%, and >50% positive cells, and the international threshold of ≥1% was used to define positive status. Routine HER2 IHC was evaluated according to the HercepTest criteria using standard local practices, with follow-up HER2 fluorescent *in situ* hybridization (FISH) as needed. Tumor grade was scored according to the Nottingham histological grade (NHG) system, which involves semiquantitative evaluation of three morphological features, tubule

formation, nuclear pleomorphism, and mitotic count, using a 3 grade scoring scheme for each feature.[2] For this study, new 4-micron slides were cut from the archival pathology blocks for ER, PgR, HER2, and Ki67 IHC, and for HER2 silver *in situ* hybridization (SISH). For NHG only the diagnostic hematoxylin and eosin slides were used. The new immunostainings were performed by a central laboratory (Helsingborg Hospital) using Ventana instrumentation: ER IHC used antibody SP1 (Ventana); PgR IHC used antibody clone 16 (Leica); HER2 SISH was performed using the INFORM Her2 Dual ISH assay (Ventana); Ki67 IHC used antibody MIB-1 (Dako). Each set of slides for each biomarker, whether the original diagnostic slides or the newly stained slides, were scored in total by three pathologists. The diagnostic slides were scored in the clinical routine, counting as the first reading, and then re-evaluated independently by two pathologists for this study (D.G. and A.E. or J.M.); the new stains were evaluated independently by all three pathologists. ER, PgR, and HER2 were evaluated as described above. Ki67 was evaluated by estimating the percentage of positive nuclei within hotspot regions, with semi-quantitative percentage scores recorded as whole numbers from 0% to 10%, then by bins of 5 from 15% to 100%. The cutoff for Ki67 was determined to be >20% high, ≤20% low, based on the internal Quality Assurance Program cutoff following the procedure recommended by the Swedish guidelines wherein one-third of cases should be high and two-thirds of cases low (see Introduction). A 'consensus score' for each biomarker was determined using majority voting from all evaluations.

For validation cohort patients, the diagnostic histopathological records for ER, PgR, HER2, Ki67, and NHG were retrieved from NKBC. For ER and PgR a cutoff for positivity of ≥1% positive cells was applied. Clinical HER2 status, positive or negative, was based on IHC and/or ISH analysis following standard guidelines. Ki67 status was based on percent positive nuclei of tumor cells as recorded in the clinical routine, with Ki67 scores thresholded at >20% being high and ≤20% as low.

Therapies

All patients were treated uniformly according to common regional guidelines that in

turn were based on national and international guidelines during the years 2006 through 2016. For the period 2010 onwards, HER2-positive patients herein generally received HER2-directed treatment with trastuzumab concomitantly with chemotherapy, with a total treatment period of 12 months for trastuzumab. ER-positive cases generally were prescribed endocrine therapy (premenopausal: 5-years tamoxifen; postmenopausal: aromatase inhibitor alone or followed by tamoxifen for a total of 5-years; extended endocrine treatment was introduced for node positive patients during the period). Most patients with ER-negative disease received chemotherapy (taxane/anthracycline), and chemotherapy for ER-positive cases was based on risk for recurrence as estimated by tumor size, nodal status, and NHG.

Tumor sample processing and RNA-sequencing

Tumor specimens were macrodissected at the pathology departments and processed in our central laboratory with handling standards that meet or exceed the recommendations of the Breast International Group, as described previously[1], with the exception that samples in the training cohort were snap-fresh-frozen instead of being preserved in RNAlater. In brief, nucleic acids were isolated using the AllPrep method and automated using QIAcube machines (Qiagen). Quality control was performed by NanoDrop spectrophotometry and BioAnalyzer (Agilent) analysis; all RNA was highly intact with RNA Integrity Number (RIN) ≥6. Starting from 1 μg total RNA, sequencing libraries for RNA-seq were generated using customized strand-specific protocols, automated for a high-throughput workflow, which have been previously described in detail.[1,3] Sequencing clusters were generated using the Illumina cBot instrument, and paired-end data were generated using an Illumina HiSeq 2000 or NextSeq 500 instrument. Sequencing statistics are presented in Supplementary Table 2.

RNA-seq gene expression measurements

Raw sequencing read data was analyzed as previously described.[1,3] To be more consistent with ongoing prospective RNA-seq data being generated within the SCAN-B initiative, for this study we truncated long sequencing reads to 2x50 bp. In brief,

raw sequencing data was demultiplexed and filtered using Bowtie 2 against ribosomal, phiX174, and UCSC RepeatMasker sequences. The remaining reads were aligned using TopHat2 2.0.5 (training cohort) or 2.0.12 and 2.0.13 (validation cohort) to the GRCh37/hg19 (with b37 masked chromosome Y and hs37d5 decoy sequences; training cohort) or the GRCh38 (validation cohort) genome together with 80,883 transcript annotations from the UCSC knownGenes table (downloaded September 10, 2012, training cohort) or 104,133 transcript annotations from the UCSC knownGenes table (downloaded September 22, 2014, validation cohort). Cufflinks v2.1.1 (training cohort) or v2.2.1 (validation cohort) was used to calculate expression levels in the form of fragments per kilobase of exon per million mapped reads (FPKM). Isoform-level gene expression data were collapsed on 27,979 (training cohort) or 30,865 (validation cohort) unique gene symbols (sum of FPKM values of each matching transcript).

Classifiers

Using only the 18,802 genes contained in the NCBI RefSeq NM category (mRNA), the gene expression FPKM values for training samples were transformed by adding a constant 0.1 to each expression value and then applying log2. Single-gene classifiers were built for the four biomarkers that have a single corresponding underlying gene by determining the optimal expression threshold for the genes *ESR1*, *PGR*, *ERBB2*, and *MKI67* that maximizes concordance with the respective histopathological consensus score within the 405 training cohort. Multi-gene classifiers for ER, PgR, HER2, Ki67, and NHG were built by training nearest shrunken centroid[4] models – using *pamr* 1.55 driven by the *caret* 6.0-47 R package – on the gene expression data of the 5000 most varying genes across all 405 samples (Supplementary Table 3), using the histopathological consensus scores for the respective biomarker as labels. In training, support vector machines (SVM) and random forests (RF) were evaluated but provided no improvement (data not shown). Model parameters were determined by performing 10 rounds of 4-fold cross-validation. During cross-validation, the 405 samples were randomly divided into four sub-cohorts, where classifiers trained on the

union of three of the sub-cohorts were used to predict the biomarker status in the remaining cohort. This procedure was repeated so that every sub-cohort was predicted by the remaining sub-cohorts exactly once, constituting one round; the result was a mean summary metric (balanced accuracy for biomarkers with dichotomous classes: ER, PgR, HER2 and Ki67; accuracy for NHG) and standard deviation (SD) for each round. The threshold yielding the round with the best mean summary metric within each biomarker (Supplementary Table 6) was used to train a prediction model using all 405 samples. The resulting nine classifiers were used to predict the IHC biomarker status of 3273 independent samples using their gene expression data normalized as described above.

The biological and functional themes of each MGC signature were evaluated using Database for Annotation, Visualization, and Integrated Discovery v6.8 (DAVID).[5] Functional annotation clustering was performed using the Entrez identifiers for each MGC signature (all genes with non-zero weight) and the default DAVID settings and annotation categories, with the following two changes to reduce the number of identified clusters: 'Classification stringency' was increased to High, and the 'Enrichment Thresholds EASE' score was decreased to 0.1.

For representational purposes, MGC classifier scores (the output of the *pamr* discriminant function) were scaled by first calculating the delta score (positive class score – negative class score) for each sample, and scaling the within-class delta scores to have a mean of 1 and -1 for the positive class and the negative class, respectively. The delta score distribution was then shifted so that a delta score < 0 represents a negative classification.

Statistical analysis

All calculations were performed using R 3.2.3. Matthews correlation coefficients (MCC) were calculated using the generalized method by Gorodkin.[6] Kappa statistics were determined using the *irr* 0.84 and *psy* 1.1 packages. Confidence intervals for kappa and MCC were calculated by bootstrapping using the *boot* 1.3 package and

10,000 bootstrap iterations. Pathology evaluations, multi-gene and single-gene predictions were compared using agreement statistics, MCC and Cohen's kappa, which were interpreted according to Viera and Garrett.[7] *P*-values ≤0.05 were considered significant.

Overall survival (OS) was used as end point for survival analysis and calculated from the date of diagnosis. Kaplan-Meier (KM) analysis and Cox proportional hazards regression were performed using the *survival* 2.38-3 package. Survival times were compared among classes using the logrank test. Multivariate Cox models included the variables age at diagnosis (continuous), lymph node status (positive vs negative), and tumor size (continuous) as covariates, as well as ER, PgR and HER2 status (all positive vs negative), and NHG (G1-G3), as relevant depending on the analyzed model (e.g., the model for the histopathologically HER2-negative group excluded HER2 status). Cases with missing data in any of the included variables were excluded from KM and Cox analysis. All models were checked for proportional hazards using Grambsch and Therneau's test for non-proportionality and Schoenfeld residuals.[8]

For calculation of concordance statistics, the following definitions are used:

Balanced Accuracy $\frac{sensitivity+specificity}{2}$

Overall Agreement $\frac{\sum_1^{NClasses}\sum SamplesTruePositiveforClassX}{\sum Samples}$

Specific Agreement for Class X (e.g. "positive" or "negative"; "G1", "G2", or "G3")

$$\frac{2 * \sum TruePositivesforClassX}{\sum PositiveReadingsforClassXbyReader1 + \sum PositiveReadingsforClassXbyReader2}$$

Expected Agreement

$$\frac{\sum_1^{NClasses}(\sum ReferenceSamplesofClassX * \sum PredictedSamplesofClassX)}{(\sum Samples)^2}$$

Kappa $\frac{OverallAgreement+ExpectedAgreement}{1-ExpectedAgreement}$

REFERENCES

1. Saal LH, Vallon-Christersson J, Häkkinen J, et al: The Sweden Cancerome Analysis Network - Breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine. Genome Medicine 7:1-12, 2015

2. Elston CW, Ellis IO: Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up. Histopathology 19:403-10, 1991

3. Häkkinen J, Nordborg N, Månsson O, et al: Implementation of an Open Source Software solution for Laboratory Information Management and automated RNAseq data analysis in a large-scale Cancer Genomics initiative using BASE with extension package Reggie. bioRxiv:1-23, 2016

4. Tibshirani R, Hastie T, Narasimhan B, et al: Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proceedings of the National Academy of Sciences of the United States of America 99:6567-6572, 2002

5. Huang da W, Sherman BT, Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44-57, 2009

6. Gorodkin J: Comparing two K-category assignments by a K-category correlation coefficient. Comput Biol Chem 28:367-74, 2004

7. Viera AJ, Garrett JM: Understanding interobserver agreement: the kappa statistic. Fam Med 37:360-3, 2005

8. Grambsch PM, Therneau TM: Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika 81:515-526, 1994