Cell-free DNA copy number variations as a marker for breast cancer in a large study cohort

Julia Beck[1] , Ekkehard Schütz[1] , Howard B Urnovitz[1], Adel Tabchy[2], William M Mitchell[3], Gordon B. Mills[2], Funda Meric-Bernstam[2]

1) Chronix Biomedical GmbH, Göttingen, Germany; 2) MDAnderson Cancer Center, Houston, TX 3) Vanderbilt University, Nashville, TN

Abstract

Background: Massive parallel sequencing provides high numbers of cell-free nucleic acid serum DNA sequences (cfDNA) that can detect trace amounts of tumor derived chromosomal imbalances and copy number variations (CNVs) in patients with cancer. The aim of this study was to determine if there is a difference between the cfDNA CNVs from patients with breast cancer (BrCa) compared to healthy controls.

Methods: DNA extracted from serum samples of 225 BrCa (Stage 1 to 4) and 205 gender and age-matched healthy controls (HC) was amplified using random primers, tagged with a unique molecular identifier per sample, sequenced on an Illumina HiSeq system and aligned to the human genome (Build 37). Hits were counted in sliding 1Mbp interval regions and normalized. Using a random-resampling procedure, a model was established to distinguish BrCa from HC using the copy number variations (CNV) and cross validated.

Results: From 1100 rounds of random resampling (50/50), a set of 31 regions was selected, based on the frequency of occurrence in the models. Using 20 random sets of a 10-fold cross validation, the selected regions were found to be highly significant discriminators between BrCa and HC ($p<10^{-5}$). When using a final linear model with 16 regions the AUC of a diagnostic ROC curve was found to be 0.895 for all samples, for Stage 1&2 the AUC was 0.86 compared to 0.93 for the higher stages. The final model included three regions from chromosome 8 and 1 and two regions from chromosome 15, the remaining regions were found as one per chromosome.

Conclusion: Using comparative massive parallel sequencing of cfDNA from cancer patients vs. controls, we were able to show that a 16-region model based on CNV, is useful to distinguish patients with breast cancer from matched controls. Genomic instabilities that are shed into the circulation from breast cancer may play a role in screening, monitoring or as companion diagnostic tests in breast cancer.

Patients and Controls

The study comprised:

I. **225 breast cancer patients** with a histopathology report of breast cancer:

223 patients had a diagnosis of invasive breast carcinoma
170 patients were judged to be at Stage I or II
46 patients were judged to be at Stage III or IV
7 patients were unstaged
2 patients had a diagnosis of ductal carcinoma in situ (DCIS)

II. **205 controls** mostly paired for age as well as other relevant characteristics and risk factors

The samples were obtained from:

MD Anderson Cancer Center (40 cases and 40 controls)
Ryazan General Hospital (50 cases and 50 controls)
Cleveland Clinics Florida (16 cases)
Proteogenex (119 cases and 112 controls)
Chronix Biomedical Germany (3 controls)

All samples were collected under an informed consent for serum collection. The study was IRB approved.

DNA Extraction, Whole Genome Amplification (WGA) and Sequencing Library Construction

Whole blood was collected and serum was recovered by centrifugation immediately after clotting. Serum samples (≥200 µL) were centrifuged at 4,000 g for 20 min to pellet any cellular debris.

Total nucleic acids were extracted from 200 µL of the supernatant with the High Pure Viral Nucleic Acid Kit (Roche Applied Science) according to the manufacturer's instructions, but without the use of carrier RNA.

Each specimen was subjected to two independent commercial whole genome amplifications (GenomePlex Whole Genome Amplification Kit, WGA, Sigma-Aldrich).

The Illumina adapter P7 necessary for sequencing and a 10 bp sample specific nucleotide sequence (index read) were added by PCR.

The barcoded DNA fragments were digested by endonuclease *Nla*III, which recognizes and cuts at each CATG sequence.

After blunting of the 5' recessed ends the library fragments were ligated to the second sequencing adapter (P5).

The resulting fragment libraries were amplified by a maximum of 10 cycles of PCR using primers complementary to the P5 and P7 adapters.

Sequencing, Mapping, and Read Counting

Sample batching: For sequencing up to 8 individuals were pooled into one sequencing library and each library was sequenced on one lane of an Illumina HiSeq1000 flowcell. A total of 57 lane-pools were prepared for the study.

Sequencing: 50 bp for each read were obtained from an HiScan SQ Sequencer (Illumina)

Mapping: The sequences were aligned to the human genome database (HG18) using the CASAVA v1.8 pipeline. All alignments with an alignment score ≥ 4 (corresponding probability of incorrect mapping is 1 in 10000) were considered as unique.

Average number of reads obtained:
6.8 M reads (SD: 1.4M) per BrCa patient
7.1 M reads (SD: 1.7M) per control

Read Counting: Hit counts in bins of 1Mbp with a 250kbp sliding window were determined using the software suite BEDtools

Data Normalization and Analyses

All data were first normalized to their total mapped read counts. To account for slide-to-slide variations, the counts per bin were normalized to the ratio per bin and slide using only samples assigned to the control group using the following equations:

1. **Number of reads per bin normalized to total number of reads**
   > for each sample individually
   > corrects for different numbers of reads between samples

$$run/slide(i): x_{n,bin} = \frac{count_{n,bin}}{\sum_{bins} count_{n,bin}}$$

where:
$count_{n,bin}$ is the number of reads per bin of an individual (n)

2. **Normalization to average normal reads per bin on flowcell in relation to all flowcells**
   > for each sample according to flowcell
   > corrects for differences between flowcells

$$\dot{Y}_{n,i,bin} = \frac{X_{n,i,bin} \times \bar{X}_{i,bin}}{\bar{X}_{all,bin}}$$

where:
$X_{n,i,bin}$ is for each bin the normalized read count of the individual (n) on slide (i)
$\bar{X}_{i,bin}$ is the average per bin over normal individuals on a slide (i)
$\bar{X}_{all,bin}$ is the average per bin over normal individuals on all slides

3) Testing for the presence of copy number variations in cfDNA – General CNI Score
For each 1000kbp bin, the normalized read counts were transferred into Z-values followed by a Parzen-Rosenblatt smoothing (X0.5, 2). For each sample, the number of bins with an absolute Z-value of >3 was determined and summed up to generate the general CNI score.

4) Cross-Validation to calculate the CNI-Region-Index
The normalized reads per bin counts were used in 1100 rounds of random resampling (50/50) yielding a set of 31 regions based on the frequency of occurrence in the models. Using 20 random sets of a 10-fold cross validation, the selected regions were found to be highly significant discriminators between BrCa and controls. For each round, the selected independent variables, their slope and the resulting AUC in the training group was recorded. The independent variables were ranked according to their occurrence and the 16 highest ranking regions were selected as final model variables. The average slope for the variables was calculated from x-validations with such co-occurrence only. Using the regions and corresponding estimates given in Table 1 the CNI-Region-Index was calculated for each sample.

6) ROC-Curves
ROC-curves were calculated for the CNI-Region-Index and the general CNI score, where AUCs were calculated using the trapezoidal rule.

Results: General cfDNA CNI Score

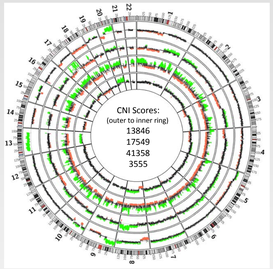

Figure 1: Circos plot of Z-values for 4 BrCa samples with **high General CNI Scores**. Z-values >2 (green), <-2 (red)

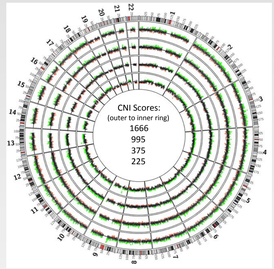

Figure 2: Circos plot of Z-values for 4 BrCa samples with **intermediate General CNI Scores**. Z-values >2 (green), <-2 (red)

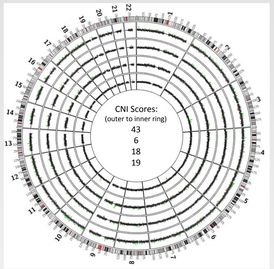

Figure 3: Circos plot of Z-values for 4 BrCa samples with **low General CNI Scores**. Z-values >2 (green), <-2 (red)

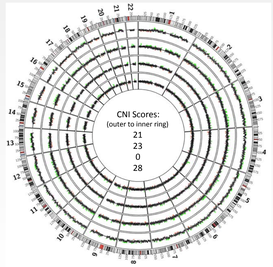

Figure 4: Circos plot of Z-Values for **4 normal controls**. Z-values >2 (green), <-2 (red)

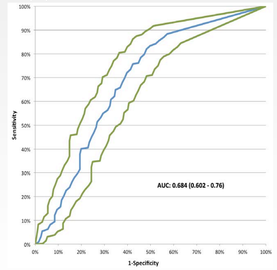

Figure 5: ROC-curve for separation of BrCa cases vs. controls by General CNI Score. Green lines represent 95% confidence interval.

The General CNI Score is based on Z-values calculated for each genomic bin (100kbp) and gives a measure of the total genomic imbalances detectable in the cfDNA.

The Circos plots of 4 patients with high and intermediate General CNI Scores (**Figure 1 and 2**) show prominent imbalances detectable in the patient's cfDNA.

On the other hand, a number of BrCa samples with low General CNI Scores are present in the study group. As exemplified for 4 such patients (**Figure 3**) the Z-value variation is not different form that seen in normal controls (**Figure 4**).

Accordingly, the resulting ROC-curve yields an AUC of only 0.68 (0.60 – 0.76) (**Figure 5**) for all samples.
AUCs for patient groups stratified by tumor stage are:
Stage 1: 0.63 (0.52 - 0.74)
Stage 2: 0.71 (0.62 - 0.80)
Stage >2: 0.69 (0.56 - 0.81)

Results: CNI-Region-Index

16 genomic bins are selected as discriminators of the MVR model (**Table 1**). The model contains several regions frequently affected by copy-number alterations in breast cancers. Examples are losses of 8p, 1q, 15q11.1 and gains of 8q, 11q14.1 and 1p13.2.

The CNI-Region-Index distributions for the different sample groups are given as boxes and whiskers plot in **Figure 6**.

The AUCs as calculated from the CNI-Region-Index indicate a high discriminative power for all BrCa patients versus all controls and for the subgroups stratified by age and tumor stage (**Figure 7 and Table 2**). The highest accuracy is achieved for higher stage tumor patients.

Table 1: The 16 regions and estimates of the multivariate regression model used to calculate the CNI-Region-Index.

Region	Cytoband	Normalization	Estimate
Losses			
chr15:19,850,001 – 20,850,000	15q11.2	Global	-0.06
chr1:142,450,001 – 143,450,000	1q21.1	Global	-0.15
chr8:2,350,001 – 3,350,000	8p23.2	Local	-0.19
chr6:118,900,001 – 119,900,000	6q22.31	Local	-0.16
chr4:173,950,001 – 174,950,000	4q34.1	Global	-0.20
chr10:115,500,001 – 116,500,000	10q25.3	Local	-0.21
chr22:20,700,001 – 21,700,000	22q11.22	Local	-0.17
chr8:11,850,001 – 12,850,000	8p23.1-22	Local	-0.13
chr5:43,250,001 – 44,250,000	5p12	Local	-0.19
Gains			
chr1:223,200,001 - 224,200,000	1q42.12	Global	0.08
chr1:113,200,001 – 114,200,000	1p13.2	Global	0.07
chr8:120,350,001 – 121,350,000	8q24.12	Global	0.20
chr9:68,250,001 – 69,250,000	9q12	Local	0.31
chr11:85,500,001 – 86,500,000	11q14.2	Local	0.20
chr15:68,600,001 – 69,600,000	15q23	Local	0.19
chrX:33,350,001 – 34,350,000	Xp21.1	Global	0.07

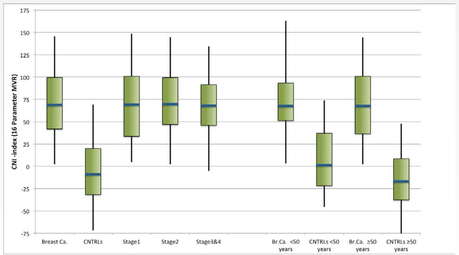

Figure 6: Boxes and whiskers plot of the different subgroups. Median (blue bar), 25th and 75th percentile (boxes) and 5th and 95th percentile.

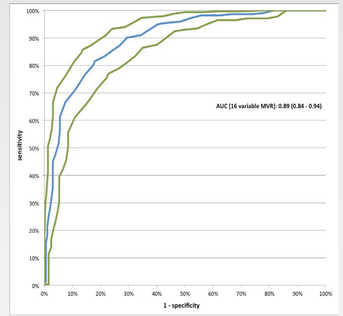

Figure 7: ROC-curve calculated from the CNI-Region-Index for all BrCa patients (222) and controls (201).

Table 2: Summary of ROC-curve calculations for patient and control subgroups stratified by age and tumor stage.

	AUC	Accuracy	Sample Number (BrCa/Cntrls)
AllVsAll	0.89 (0.84 - 0.94)	0.82 (0.77 - 0.86)	222/201
Age < 50	0.86 (0.76 - 0.94)	0.80 (0.72 - 0.87)	95/75
Age ≥ 50	0.92 (0.85 - 0.96)	0.84 (0.78 - 0.89)	127/126
Stage 1	0.89 (0.81 - 0.94)	0.85 (0.81 - 0.90)	61/201
Stage 2	0.90 (0.83 - 0.95)	0.85 (0.80 - 0.89)	109/201
Stage 3&4	0.88 (0.79 - 0.95)	0.89 (0.84 - 0.93)	46/201

Summary and Conclusions

- **Using comparative massive parallel sequencing of cfDNA a 16-region model based on copy-number imbalances was developed.**
- **The CNI-Region-Index calculated from this model has a higher discriminative power than the General cfDNA CNI Score that measures the overall genomic instability.**
- **Breast cancer can be detected with an accuracy of 0.82 over the whole dataset in this study.**