DNA copy number profiles during tumour progression in breast cancer

Hiroko Kato Solvang^{1*}, Hidetoshi Shimodaira², Fatemeh Kaveh³, Daniel Nebdal⁴, Anne-Lise Børresen-Dale^{4,5}, Hege Edvardsen⁶ and Vessela N Kristensen^{4,5,6}

*Correspondence: hiroko.solvang@imr.no

CrossMark

← Click for updates

¹Marine Mammals Research Group, Institute of Marine Research, Bergen, Norway.

²Division of Mathematical Science, Graduate School of Engineering Science, Osaka University, Japan.

³Medical Genetics Department, Oslo University Hospital (Ullevål), Kirkeveien 166, 0407 Oslo, Norway.

⁴Department of Genetics, Institute for Cancer Research, Oslo University Hospital, Montebello, 0310 Oslo, Norway.

⁵Institute for Clinical Medicine, Faculty of Medicine, University of Oslo, Norway.

⁶Department of Clinical Molecular Biology and Laboratory Science (EpiGen), Division of Medicine, Akershus Universit Hospital, University of Oslo, Norway.

Abstract

Background: In this paper we report the prevalence of copy number aberration events at various stages (subclasses) of breast cancer as assessed by two different statistical methods, GISTIC, a well-known method to identify significant somatic copy-number alterations in a given set and GSDIP: Genomic Identification of Significant Differences in Progression, a newly developed numerical algorithm described here to compare different sets.

Methods: GISTIC assesses significant aberrations towards whole genomic location for each stage and GSDIP assesses significant difference towards whole genomic location for two stages pair-wise.

Results: We compare GSDIP with the original GISTIC by a simulation study. Comparing with the focal/broad regions obtained by GISTIC, GSDIP can directly illustrate significantly different regions between different stages. We then performed experimental data analyses and identified all significant aberrations at each stage of breast cancer using GISTIC and identified the peaks significant for each stage/tumour category. Then we applied GSDIP to identify the significant differences between the tumour size categories in progression. The significance was assessed by permuting the data across all samples in all stages. The statistically significant copy number aberrations specific for the various stages are presented here by both statistics on a set of 530 cases (validation set). The same analysis was performed in subsets of cases according to ER-status. The genes located in the identified regions were investigated by applying network analysis and the list for genomic loci related to allele-specific disparity.

Conclusions: Our method indicates significant particular regions related to difference of copy number aberration between different stages. The particular biological characteristics for network analysis on estimated regions are summarized.

Keywords: DNA copy number alteration, breast cancer, computational algorithm, permutation, false discovery rate, allele-specific disparity

Introduction

Breast cancer is heterogeneous disease both at clinical and molecular level. Molecular analyses of cancer cells in various stages (subclasses) of progression have revealed that alterations in tumour suppressor genes and oncogenes accumulate during tumour progression and correlate with the

clinical aggressiveness of cancer. Tumour size, indicated by the T-category, is known as a strong prognostic indicator for breast cancer and is one of the factors taken into account when deciding how and whether to treat a patient, independently of lymph node status. Although imperfect, tumour size is also together with grade a surrogate marker of time and

progression. Gene expression profiling T1 and T2 tumours was recently attempted and revealed distinct molecular pathways characterizing each tumour category [3,4]. The aim of the present study is to deepen our analysis at array CGH DNA level and compare profiles of various stages of breast tumour as a surrogate of the course of tumorigenesis and identify significant aberrations that may be driving the process at each stage.

To identify somatic copy-number alterations (SCNA) related to tumour development and progression methods, such as Genomic Identification of significant targets in cancer (GISTIC) [1] were applied. This method identifies broad/focal regions of the genome that are significantly amplified or deleted across a set of samples. These regions are recognized as likely drivers SCNAs by evaluating the frequency and amplitude of the observed events. The main procedures to identify the regions take the summation for the aberrations for tumours within a certain class and find significantly aberrant regions using permuted data. GISTIC II [2] adds to the original GISTIC methodological improvements for different complex cancer genomes containing a mixture of SCNA types occurring at distinct background rate. The outputs are very sensitive to sample size. In fact, GISTIC II indicated that most significant peaks become stable around a set as large as 800 samples [2].

Our aim was to identify significant differences between the copy number alteration landscapes in different time points of tumour progression. For this purpose it is not enough to identify SCNA for each stage of progression, but also to assess the significance of difference in their occurrence between the stages. Applying GISTIC for different stages require two steps: 1) apply GISTIC to each stage and 2) compare the significant focal regions between stages. Furthermore, the number of samples for each stage usually varies and the output assessed by permutation within one stage could not be compared with other stages.

Therefore, for the analysis of data involving inconsistent sample number for each stage, we propose a new approach called 'Genomic/identification of Significant Difference for DNA copy number profiles/n tumour Progression (GISDIP)', which directly identifies the differences of SCNAs between different stages through one step procedure. GISDIP uses all relevant samples to generate the null distribution to identify the level of significance.

In this study, we apply GISDIP to CGH DNA profiles for the Norwegian Breast Cancer Study consisting of 530 samples from five different cohorts, stratified by T-status and ER-status. The significant regions obtained by GISDIP were subjected to additional analyses such as network function analysis of the harboring genes and allelic-specific disparity [5] that detects allele-specific chromosomal aberrations and loss of heterozygosity.

In the following sections we describe the principle of GISDIP and compare the results obtained by GISTIC and GISDIP as well as their performance using a simulation study.

Methods

Segmentation for CGH data

In this study, array CGH DNA profiles were observed as log R (denoted hereafter as logR), which is the log-ratio of observed probe intensity to expected intensity. Positive/negative logR indicate a region of DNA copy number gain/loss. Every tumour genome was segmented by fitting a piecewise constant function to the logR values to obtain smoothing value [6].

GISTIC

Genomic Identification for Significant Targets (GISTIC) was developed to identify broad/focal regions of the genome that are significantly amplified or deleted across a set of samples. Each aberration is assigned a G-score that considers the amplitude of the aberration as well as the frequency of its occurrence across samples. False Discovery Rate q-value [7] are then calculated for the aberrant regions, and regions with q-values below a user-defined threshold (default=0.25) are considered significant. Identified 'peak region' is the part of the aberrant region with greatest amplitude and frequency of alteration. The original algorithm was published in 2007 [1] and it has revised as GISTIC II [2] that additionally describes a probabilistic method for defining the boundaries of selected SCNA regions with user-defined confidence. Since each aberration is considered across samples, resolution for average number of genes per peak becomes sensitive for sample size. In article [2] indicated that stable resolution for number of focal peaks and genes per peak can be obtained at sample sizes over 800-1000 samples.

GISDIP

For given CGH data (logR values), let the smoothed value indicate x_{ijk} for genomic location $i=1,2,\dots,N$, stage $j=1,2,\dots,L$ and samples $k=1,2,\dots,m_j$ for stage j . First, the mean value y_{ij} of x_{ijk} is calculated as a representative value for each genomic location of each stage defined by

$$y_{ij} = \frac{1}{m_j} \sum_{k=1}^{m_j} x_{ijk} \quad (1)$$

Next, a statistics S_i is calculated to investigate difference of y_{ij} between different stages. If the difference between a stage indicating the largest mean and a stage (e.g., $j=1$) indicating the smallest mean, S_i should be defined as. $S_i = \max_{j=1} y_{ij} - \min_{j=1} y_{ij}$. On the other hand, if each difference between two different stages S_i should be taken into account, should be defined as $S_i = |y_{i1} - y_{i2}|$ or $S_i = y_{i1} - y_{i2}$. The later formula detect $y_{i2} > y_{i1}$ in addition to the difference between two stages. To avoid the variability for each stage, S_i should be standardized dividing by the variance σ_i for each genomic location of all stages that

$$\sigma_i = \sqrt{\frac{1}{(m_1 + m_2 + \dots + m_L)} \sum_{j=1}^L \sum_{k=m_1}^{m_j} (x_{ijk} - \bar{x}_i)^2} \quad (2)$$

Now, to assess S_{ij} obtained by original data, GISDIP generates the null distribution for S_i by the following procedures:

1. Aggregating all samples for all stages and labeling the same number on samples for same stage (if there are four stages, 1–4 are labeled on the samples).
2. Permuting the number for each stage across whole samples.
3. Taking the same number for each stage from permuted data and assign them to each stage.
4. Calculate (1) and obtain S_i^* .

If 1000 times iteration is performed through the above procedures 2-4, we can generate a null distribution by $S_i^{*(1)}, \dots, S_i^{*(1000)}$. In this case, the p-value for each genomic location is calculated by

$$p_i = \frac{\# \{S_i^{*(b)} \geq S_i, b = 1, 2, \dots, 1000\}}{1000}. \quad (3)$$

To assess multiple testing, the false discovery rate (FDR [7]) should be applied to these p-values. The regions with $FDR < 5\%$ are taken to indicate significant difference between different stages in this study. The detail of the above procedures is summarized in **Figure 1**.

Allele-specific disparity

SNP-based arrays can detect allele-specific chromosomal aberrations such as loss of heterozygosity. We and others have developed several novel approaches to identify loci showing preferential disparity from heterozygous toward either the A/B-allele homozygous (allelic-disparity, AD) [5]. In a previous study we applied the method to B-allele frequency and logR for blood and tumour of breast cancer patients and concluded that 13 K loci were subjected to AD and associated with some clinical parameters like grade, ER, Her2, TP53 and PR. In this article, we analyzed the stage specific markers obtained by GISDIP for allelic disparity by overlapping with the 13K loci obtained in [5].

Pathway analysis

Pathway analysis was applied in order to investigate the network function, canonical pathway and biological functional interaction for the lists of markers obtained by GISDIP. This analysis was performed by Ingenuity Pathway Analysis (IPA, Ingenuity system <http://www.ingenuity.com/>). This delivers an assessment of the signaling and metabolic pathway, molecular networks and biological process that are most significantly perturbed in the dataset of interest. IPA offers many options for finding insights into relationships, mechanisms, function and pathway of relevance.

A simulation study

Before applying both methods to real data, we performed a simulation study. Suppose $u = \{(x_{ij}, y_{ij}), i = 1, \dots, m, j = 1, \dots, n\}$ be the copy number data where m is the number of samples

Figure 1. Procedures to generate permuted data.

Let x_{ijk} be the smoothed values of CGH data (logR) for the genomic location $i = 1, 2, \dots, N$, stage $j = 1, 2, \dots, L$ and samples $k = 1, 2, \dots, m_j$ for stage j . Mean value y_{ij} of x_{ijk} is calculated for each genomic location of each stage. First, the statistics S_i is calculated to investigate difference of y_{ij} between different stages: $S_i = \max y_{ij} - \min y_{ij}$ where $\max y_{ij}$ indicates the largest mean and $\min y_{ij}$ point to the smallest mean. Then each difference between two different stages takes into account by $S_{ij} = |y_{i1} - y_{i2}|$ or $S_{ij} = y_{i1} - y_{i2}$. The later formula detect in addition to the difference between two stages. To avoid the variability and standardization of each stage S_i is divided by the variance σ_i for each genomic location. To assess S_i (obtained by original data) GISDIP generates H_0 by: 1) Aggregating all samples of all stages and labeling. 2) Permuting the numbers. 3) Picking up the same amount of samples for each stage and redefine the data as x_{ijk}^* . 4) Calculate y_{ij}^* by S_i^* . 5) Iterate B times to generate the H_0 by $S_i^{*(1)}, \dots, S_i^{*(B)}$.

$y_{i1}, y_{i2}, \dots, y_{im}$ are the log-ratio measurements of one sample that are ordered by the location of a probe along one chromosome ($x_{i1} \leq x_{i2} \leq \dots \leq x_{im}$) where n is the probe number. In the first set of simulation, the minimum length of probes was equal to 50 and the maximum length of the probes was considered 700 probes. We generated the probes with the normal distribution and the standard deviation of 0.10. To emulate the complexity of real tumour profiles, we added aberrations by uniform distribution that generates random deviates or aberrations. Any values less than 1 were set as a deletion and values greater than two was fixed as amplification. **Supplement figure S1** illustrated 25 examples of generated data for each case. Simulation 1 (sim1) involved 30 samples and 1000 artificial features with four amplified and two deleted artificial aberrations. Simulation 2 (sim2) involved 100 samples and 1000 artificial features with six amplifications and four deletions as artificial aberrations. To the two simulation data, we applied the basic theory for GISTIC and GISDIP. We did not apply the entire procedure of GISTIC to enhance the essential difference between GISTIC and GISDIP. In the case of GISTIC (see **Figure 1** in [1]), the program first calculates summation [1] (called G-score, $-\log(\text{probability} | \text{background})$) in the case of

GISTIC II [2]) across all samples of a data. To assess the vector of G-score, permutation across the features was performed. For 1000 permuted data, the null distribution was generated for each marker across the genome. As final procedure, the program assesses G-scores for the null distribution and applies q-value to the vector of obtained p-values. In this study, FDR was calculated instead of q-value. GISTIC also involved peel-off algorithm to rescore to potential peaks; however, we here apply only basic procedure (not applying entire procedure of GISTIC) as below:

1. Calculating G-score for sim1 and sim2 across the samples.
2. To assess G-score obtained by original data, generate null-distribution by permuting features and calculating G-score again.
3. Calculating p-value for G-score to null-distribution and apply FDR.

Supplement figure S2 summarized the outputs for GISTIC. The solid black plots for two panels indicate G-score for sim1 and sim2. The significant aberrations obtained by GISTIC were indicated by circles for each panel. The regions indicated the highest peaks for amplified and deleted aberrations. On the other hand, the significant regions obtained by GSDIP were illustrated by lines the lowest panels in **Supplement figure S3**. The top and middle panels of **Supplement figure S3** indicated mean values for sim1 and sim2. The curve of the lowest panel of **Supplement figure S3** was the absolute value of the difference between mean values of sim1 and sim2. As shown in the lowest panels, the lines obtained by GSDIP indicated clear difference between mean values of sim1 and sim2. When applying GISTIC to this curve, the region marked by circle was estimated as significant peak region. These studies show that while similar profiles were identified by GISTIC and GSDIP, GSDIP had a higher sensitivity to identify significant peaks in smaller subsets. Therefore, GSDIP should be appropriate to identify the significant difference for DNA copy number profiles in the stages for tumour progress.

Materials

We applied GISTIC and GSDIP to logR values transferred from DNA copy number data, which high resolution CGH profiling was carried out by Agilent technologies on 244K Agilent Human Genome CGH Microarrays. These data came from the Norwegian Breast Cancer Study comprising of five different cohorts, designated as MDG, Uppsala, Ull, MicMa and DBCG. MDG [8] consists of pure invasive ductal carcinomas (DCIS) less than 15 mm and mixed invasive and *in situ* components, stage I (T1, <2cm), stage II (T2, 2cm ≤ and <5cm), and stage III (T3, 5cm ≤). Uppsala [8] came from core biopsies of women with dense breast parenchyma and invasive ductal carcinomas. Both cohorts are clinically similar. Ull [9] consists of mainly stage I and II breast cancers. MicMa consists of mainly stage I and II breast cancers. Fresh frozen tumour biopsies were collected from 150 to 920 patients included in the Oslo Micro-metastasis Project between 1995 and 1998. DBCG comprise

a collection of tumour tissues from 3,083 high-risk Danish breast cancer patients diagnosed in the period 1982-1990 [10-12]. Number of patients for each stage in five cohorts is summarized in **Table 1A**.

The 530 cases analyzed here have been characterized for all routine clinical parameters. We have in this study performed the analysis on the entire data set as well as stratified by estrogen receptor (ER), which is well-known important parameter in breast cancer [13]. The samples for each stage are shown in **Table 1B**. A few patients of each stage were not observed ER status. For the reason, the total number of ER status of each stage was less than total number of each stage. We applied GSDIP to all samples shown in **Table 1A** as well as stratified by ER as shown in **Table 1B**.

Table 1. Summary of the data for five cohorts and T-status.

Table 1A					
Patient material	Stage/subclass				Total
	DCIS	T1	T2	T3	
MDG	2	21	19	1	43
Uppsala	27	54	11	0	92
Ull	1	61	92	13	167
MicMa	0	22	19	4	45
DBCG	0	57	100	26	183
Total	30	215	241	44	530

Table 1B		
Stages	ER	
	Positive (+)	Negative (-)
DCIS (30)	22	6
T1 (215)	154	51
T2 (241)	168	70
T3 (44)	22	19
Total	366	146

Results and discussion

Applying GISTIC to all samples of different stages

GISTIC was first used for the analysis of all samples for different stages and tumour categories. Two plots of G-scores and q-values were obtained with respect to amplifications and deletions separately and are shown for all markers by red and blue lines in all four stages in **Figure 2**. For detecting significant gain and loss, we used the default threshold 0.25. The symbol '*' and 'x' indicate the areas corresponding to the areas identified also by GSDIP as significant for each stage (see below). The exact *amp* and *del* loci and residing genes identified by this method are summarized in **Supplement Tables S1a-d**. Gene rich loci amplified at 1q32.1, 8q23, 20q13.33, and deletions in multiple others (**Supplement Table S1a**) were observed at DCIS level. However many of the same loci continued to be

significant by GISTIC in the next stages (**Supplement Tables S1b-d**). Therefore a formal method to compare the differences between the significant peaks at each stage was needed.

Applying GISDIP to all samples for different stages

GISDIP was applied to compare pair wise each two neighboring stages (subclasses) as follows DCIS and T1 (DCIS-T1), T1 and T2 (T1-T2), and T2 and T3 (T2-T3). The absolute value $S_{ij} = |y_{iA} - y_{iB}|$ where A and B are two neighbor stages and the p-values calculated by (3) were summarized in **Figure 3**. The plots for the significant regions illustrate negative/positive difference $S_{ij} = y_{iA} - y_{iB}$ between stages A and B (for example) instead of the absolute amount calculated by $S_{ij} = |y_{iA} - y_{iB}|$. For each difference $S_{ij} = y_{iDCIS} - y_{iT1}$ for DCIS-T1, $S_{ij} = y_{iT1} - y_{iT2}$ for

T1-T2, and $S_{ij} = y_{iT2} - y_{iT3}$ for T2-T3 were calculated. The positive plots (above zero) mean that the aberrations of stage A are larger than the aberrations of stage B, and the negative plots (below zero) mean that the aberrations of stage B are larger than the aberrations of stage A, e.g., for 'DCIS-T1', stage A corresponds to 'DCIS' and stage B corresponds to 'T1'. The significant regions by 5% FDR between DCIS and T1, T1 and T2, and T2 and T3 are also summarized in **Supplement Table S2**.

The total number of markers (genes) estimated as significant different regions (FDR 5%) by GISDIP were 1923 (99) when comparing DCIS to T1, 5558 (626), comparing T1-T2, and 1035 (53) comparing T2-T3, respectively. The probes and corresponding gene symbols are summarized in **Supplement Table S2**. Comparing our data with the TCGA data on breast cancer ([14],

Figure 2. Significant aberration (amplification/deleted) estimated by GISTIC.

Significant aberrations estimated by GISTIC in cases of DCIS, T1, T2 and T3. Left and right panels indicate amplification (red color) and deletion (blue color). The symbols •, * and × indicates significant areas for stages: DCIS-T1, T1-T2, and T2-T3 obtained by GISDIP.

we see that among the focal amplifications/deletions and arm-level gains and losses identified by the TCGA network we are able to suggest a time-line to these CAN events ([Table 2](#)). Applying GISDIP a more clarified picture of the stage dependent events emerges. Only the deletion on chromosome 3q21.3, and amplification of 16q12.1 and 23 are significantly predominant in DCIS, while chromosomes 1,5,6,8,13,14,16,19 and 20, are characteristic of stage I, and those on 7,9 and 11-for stage 3 ([Figure 3](#)).

Focusing on the overlapping regions identified by both GISTIC and GISDIP we find for DCIS only the deletion on 3q21.3 and amplification of 16q12.1 remaining. For the T1-T2 comparison a main finding is the increasing frequency of amplifications on 1q (regions 1q21.2, 1q21.1 and 1q31 among others) and 8q24.3 and 9q34.3. For T2-T3 we see, among others, an increasing frequency of amplification of 11q13.4-5 as well as deletion of 7q35 only found in T3 ([Supplement Table S3](#)). Among the gene lists from the identified regions, we found in

Table 2. Comparison of GISTIC results with the TCGA data on breast cancer.

A - Overlapping GISTIC - DCIS			
Amplification		Deletion	
Cytoband	Genes	Cytoband	Genes
8p11.23	No overlapping genes	3p21.31	C3orf18, C3orf45, CACNA2D2, CYB561D2, GNAI2, GNAT1, HEMK1, hsa-mir-566, HYAL1, HYAL2, HYAL3, IFRD2, NAT6, RASSF1, SEMA3B, SEMA3F, SLC38A3, TMEM115, TUSC2, ZMYND10
11q13.3	No overlapping genes	4p16.3	ATP5I, CPLX1, DGKQ, FGFR1, GAK, hsa-mir-571, IDUA, MFSD7, MYL5, PCGF3, PDE6B, PIGG, RNF212, SLC26A1, SPON2, TMEM175, ZNF141, ZNF595, ZNF718, ZNF721
16q12.1	No overlapping genes	7p22.3	C7orf50, COX19, CYP2W1, FAM20C, GPER, GPR146, HEATR2, hsa-mir-339, INTS1, MAD1L1, MAFK, MICALL2, PDGFA, PRKAR1B, PSMG3, TMEM184A, UNCX, ZFAND2A
17q12	ERBB2	8p23.2	CSMD1
		8q24.3	ADCK5, BOP1, C8orf73, CPSF1, CYC1, CYHR1, DGAT1, EEF1D, EPPK1, EXOSC4, FAM83H, FBXL6, FOXH1, GLI4, GPAA1, GPIHBP1, GPR172A, GPT, GRINA, hsa-mir-661, hsa-mir-939, HSF1, KIFC2, LRRC14, LRRC24, LY6H, MAF1
		9p13.1	ANKRD20A3, CNTNAP3, FAM74A3, FAM75A2, FAM75A3, FAM75A5, FOXD4L2, ZNF658
		11p15.5	B4GALNT4, C11orf35, HRAS, hsa-mir-210, IRF7, LRRC56, PKP3, PTDSS2, RASSF7, RNH1, SCT, SIGIRR
		11q13.1	No overlapping genes
		12q24.33	GALNT9
		15q11.2	OR4N4, OR4M2
		19p13.3	ATP5D, C19orf23, C19orf24, C19orf26, CIRBP, EFNA2, GPX4, MIDN, MUM1, SBNO2, STK11
B - Overlapping GISTIC - T1			
Amplification		Deletion	
Cytoband	Genes	Cytoband	Genes
1q21.3	No overlapping genes	2q37.3	AGXT, ANKMY1, AQP12A, AQP12B, ATG4B, BOK, C2orf54, CAPN10, D2HGDH, DTYMK, DUSP28, FARP2, GAL3ST2, GPC1, GPR35, HDLBP, hsa-mir-149, ING5, KIF1A, MTERFD2, NEU4, PASK, PDCC1, PPP1R7, RNPEPL1
1q44	OR2T27, PGBD2, SH3BP5L, ZNF672, ZNF692	3p12.3	C3orf38, CADM2, CGGBP1, CHMP2B, GBE1, HTR1F, POU1F1, ROBO1, ROBO2, VGLL3, ZNF654
8p11.23	ZNF703	3p21.31	C3orf45, CACNA2D2, CYB561D2, GNAI2, HYAL1, HYAL2, HYAL3, IFRD2, NAT6, RASSF1, SEMA3B, SLC38A3, TMEM115, TUSC2, ZMYND10
8q24.21	MYC	4p16.3	PDE6B
11q14.1	C11orf67, CLNS1A, RSF1	7p22.3	C7orf50, COX19, FAM20C, CYP2W1, GPER, GPR146, HEATR2, hsa-mir-339, INTS1, MAD1L1, MAFK, MICALL2, PDGFA, PRKAR1B, PSMG3, TMEM184A, UNCX, ZFAND2A
12q15	CPM, MDM2, CPSF6	8p21.2	ADAM7, ADAM28, ADAMDEC1, STC1
17q12	ERBB2	10q23.31	PTEN
17q23.1	TUBD1, RPS6KB1	10q26.3	ADAM8, GPR123, hsa-mir-202, KNDCC1, TUBGCP2, UTF1, VENTX
20q13.2	ZNF217	11p15.5	AP2A2, ATHL1, B4GALNT4, BET1L, BRSK2, C11orf35, CD151, CEND1, CHID1, DEAF1, DRD4, EFCAB4A, EPS8L2, HRAS, hsa-mir-210, IFITM1, IFITM2, IFITM3, IFITM5, IRF7, LRDD, LRRC56, MUC2, MUC5B, MUC6, NLRP6, ODF3, PDDC1, PKP3, PNPLA2, POLR2L, PSMD13, PTDSS2, RASSF7, RIC8A, RNH1, RPLP2, SCGB1C1, SCT, SIGIRR, SIRT3, SLC25A22, TALDO1, TMEM80, TOLLIP, TSPAN4
		12q24.33	GALNT9
		15q11.2	OR4M2

Continuation of Table 2.

B - Overlapping GISTIC - T1

Amplification		Deletion	
Cytoband	Genes	Cytoband	Genes
		17p12	hsa-mir-744, MAP2K4
		19p13.3	ATP5D, C19orf23, C19orf24, C19orf26, CIRBP, EFNA2, GPX4, MIDN, MUM1, SBNO2, STK11

C - Overlapping GISTIC - T2

Amplification		Deletion	
Cytoband	Genes	Cytoband	Genes
1q44	LOC646627, OR2T27, PGBD2, SH3BP5L, ZNF672, ZNF692	2q37.3	AGXT, ANKMY1, AQP12A, AQP12B, ASB1, ATG4B, BOK, C2orf54, CAPN10, D2HGDH, DTYMK, DUSP28, ESPNL, FARP2, GAL3ST2, GPC1, GPR35, HDAC4, HDLBP, HES6, hsa-mir-149, ILKAP, ING5, KIF1A, KLHL30, MTERFD2
20q13.2	ZNF217	3p12.3	GBE1, ROBO1, ROBO2
7p11.2	EGFR, LANCL2	6q27	C6orf124, DACT2, FRMD1, KIF25, MLLT4, SMOC2
8p11.21	No overlapping genes	8p21.2	ADAM28, ADAM7, ADAMDEC1, NKX3-1, STC1
8p11.23	No overlapping genes	8p23.2	CSMD1
8q24.21	MYC	9p21.3	CDKN2A, CDKN2B, MTAP
11p13	C11orf41, CCDC73, CD59, CSTF3, DEPDC7, EIF3M, FBXO3, HIPK3, PRRG4, QSER1, TCP11L1, WIT1, WT1	9p24.1	PTPRD
11q14.1	C11orf67, INTS4, RSF1	10q23.31	PTEN
12p13.33	IQSEC3, SLC6A12, SLC6A13	13q14.2	RB1, LPAR6
12q15	CPSF6, FRS2, LYZ, YEATS4	15q11.2	OR4M2
17q12	ERBB2	17p12	hsa-mir-744, MAP2K4
17q23.1	TUBD1, RPS6KB1	19p13.3	No overlapping genes
19q12	CCNE1		

D - Overlapping GISTIC - T3

Amplification		Deletion	
Cytoband	Genes	Cytoband	Genes
1q44	LOC646627, OR2T27, PGBD2, SH3BP5L,	3p21.31	ABHD14A, ABHD14B, ACY1, ALAS1, AMIGO3, AMT, APEH, ARIH2, ATRIP, BAP1, BSN, C3orf18, C3orf45, C3orf54, C3orf62, CACNA2D2, CAMKV, CAMP, CCDC36, CCDC51, CCDC71, CCDC72, CDC25A, CELSR3, CISH, COL7A1, CYB561D2, DAG1, DALRD3, DCP1A, DNAH1, DOCK3, DUSP7, FBXW12, GLT8D1, GLYCTK, GMPPB, GNAI2, GNAT1, GNL3, GPR62, GPX1, GRM2, HEMK1, hsa-let-7g, hsa-mir-135a-1, hsa-mir-191, hsa-mir-566, HYAL1, HYAL2, HYAL3, IFRD2, IMPDH2, IQCF1, IQCF2, IQCF3, ITIH1, ITIH3, ITIH4, KLHDC8B, LAMB2, LOC646498, MAPKAPK3, MON1A, MST1, MST1R, MUSTN1, NAT6, NCKIPSD, NEK4, NICN1, NISCH, NME6, NT5DC2, PARP3, PBRM1, PCBP4, PFKFB4, PHF7, PLXNB1, PPM1M, PRKAR2A, PRKCD, QARS, QRIH1, RAD54L2, RASSF1, RBM15B, RBM5, RBM6, RFT1, RHOA, RNF123, RPL29, RRP9, SEMA3B, SEMA3F, SEMA3G, SFMBT1, SLC25A20, SLC26A6, SLC38A3, SPCS1, SPINK8, STAB1, TCTA, TEX264, TKT, TLR9, TMEM110, TMEM115, TMEM89, TNNC1, TRAIP, TREX1, TUSC2, TWF2, UBA7, UCN2, UQCRC1, USP19, USP4, WDR6, WDR82, VPRBP, ZMYND10
6q21	hsa-mir-587		FOXO3, ARMC2, LACE1
8p11.23	No overlapping genes		C6orf124, DACT2, FRMD1, GPR31, KIF25, MLLT4, SMOC2, TCP10, TTLL2, UNC93A
8q24.21	MYC		CSMD1

Continuation of Table 2.

D - Overlapping GISTIC - T3			
Amplification		Deletion	
Cytoband	Genes	Cytoband	Genes
10q22.3	No overlapping genes		ADAM8, BNIP3, C10orf125, C10orf91, C10orf93, CALY, CLRN3, CYP2E1, DPYSL4, EBF3, ECHS1, FOXI2, FRG2B, GLRX3, GPR123, hsa-mir-202, INPP5A, JAKMIP3, KNDC1, LRRC27, MGMT, MKI67, MTG1, NKX6-2, NPS, PAOX, PPP2R2D, PRAP1, PTPRE, PWWP2B, SPRN, STK32C, SYCE1, TCERG1L, TUBGCP2, UTF1, VENTX, ZNF511
11p13	CCDC73, PRRG4		hsa-mir-744, MAP2K4
11q14.1	C11orf67, INTS4, RSF1		
12p13.33	IQSEC3, SLC6A12, SLC6A13		
17q12	ERBB2		

DCIS-T1 deleted cluster of differentiation markers CD80 and CD86 providing co-stimulatory signals necessary for T-cell activation [15,16]; in T1-T2 we find in RBBP5, a protein coding and tumour suppressor, PIK3C2B (phosphatidylinositol 3 kinase pathway) and MDM4 for p53 inhibitor like MDM2 [17]; in T2-T3 we see specifically deletion of the tumour suppressor genes CDKN2A and CDKN2B (Figure 4).

Pathway analysis for GSDIP outputs to all samples for different stages

We further examined whether the genes identified by GSDIP as differentially amplified or deleted between pre-invasive and invasive tumours of different size categories are enriched for certain functions, we applied IPA as introduced in sub-section Method-Pathway analysis to these markers. Using IPA we investigated the canonical pathways and network functions enriched within the GSDIP gene list. As shown in 'Top canonical pathway' of Table 3, the events specific for DCIS included genes from DNA Double-Strand Break Repair by Homologous Recombination, B Cell Development, PPARα/RXRα Activation, Autoimmune Thyroid Disease Signaling and Allograft Rejection Signaling. The molecules indicated ATRX, POLA1, IL1RAPL1, and GK located on chromosome X. For T1-T2, all canonical pathways were related to regions indicating specific to T1 aberrations when compared to T2 aberrations. For differences between T2-T3, most canonical pathways contain molecules of CDK2A on chromosome 9, FZD4 on chromosome 11 for T2 aberrations > T3 aberrations, and WNT11 on chromosome 11 for T3 aberrations > T2 aberrations.

ER stratification for different stages

In addition, we applied GSDIP to the samples stratified by ER status (positive/negative). Estimated significant regions were summarized in Supplement Table S4. In the ER negative cases no significant regions were identified by GSDIP when comparing DCIS-T1 and T1-T2. Comparing T2-T3 however we found several loci significantly different between the two

subclasses such as 1p31.1, 6q14.1 and 9p21-22 (Figure 5). When overlapping these data with the GISTIC data only the region at 9p21 was found by both methods (Figure 6 and Supplement Table S4). Much more distinct differences were observed in the ER positive cases. The estimated number of markers was 273 for T2-T3 in ER negative and 474 for DCIS-T1, 6020 for T1-T2, 517 for T2-T3 in ER positive (Supplement Table S4, Figure 5). The overlapping results from GISTIC and GSDIP analyses indicate significant copy number alterations in ER+ (T1-T2) for regions such as 2q11, 4p16, 8q24, 11q13, 16p13 and 17q25 (Figure 6 and Supplement Table S4). In general, it is known that copy number alterations are better defined in ER positive cases than copy number alternations for ER negative, where a multiple firestorms events are observed [18]. This fact and the smaller number of cases may support that it was hard to have significant differences between DCIS-T1 and T1-T2.

We applied IPA to the markers obtained by GSDIP applying to ER negative samples for T2-T3 and positive samples for all cases. IPA investigated the canonical pathway and network function. The outputs were summarized in Table 4.

Overlapped regions with allele-specific disparity (AD)

As introduced in sub-section Method-GSDIP, in one of the datasets (MicMa) we had available SNP-CGH data and found 13K loci in which we observed a preferential loss or gain of a specific allele (referred here to as allelic disparity (AD)). Illumina Beadstudio package was applied to MicMa samples and handled B-allele frequency and logR values as described in Kaveh et al [5]. We matched the identified aberrant list on RefSeq gene IDs from this study (Agilent 44K array) with the positions of the rs-numbers for SNPs, using an annotation file called "SNP_annotation_Infinium HumanBeadChip1_v1.2.1_allchr" released by Illumina. In 13K loci, the overlapped number and corresponding chromosome regions were summarized in the cases of all samples and samples for ER positive in Table 5. According to three combinations of different stages, DCIS-T1, T1-T2, and T2-T3, the probabilities for the events of allelic disparity

Table 3. IPA outputs of top canonical pathways and top associated network functions for different stagesa.

Top canonical pathway		-log (B-H p-value)
DCIS-T1	DNA Double-Strand Break Repair by Homologous Recombination	0.481
	B Cell Development	0.405
	PPARα/RXRα Activation	0.405
	Autoimmune Thyroid Disease Signaling	0.405
	Allograft Rejection Signaling	0.405
T1-T2	Glutathione Metabolism	1.200
	VEGF Signaling	0.330
	mTOR Signaling	0.330
	Hypoxia Signaling in the Cardiovascular System	0.330
	Small Cell Lung Cancer Signaling	1.320
T2-T3	Ovarian Cancer Signaling	1.270
	Factors Promoting Cardiogenesis in Vertebrates	1.010
	Glioblastoma Multiforme Signaling	0.708
	Cell Cycle: G1/S Checkpoint Regulation	0.708
	Wnt/β-catenin Signaling	0.0179
Top associated network function		Score
DCIS-T1	Hematological System development and Function, Tissue Morphology, Cellular Growth and Proliferation	35
	Genetic Disorder, Reproductive System Disease, Developmental Disorder	33
	Cancer, Dermatological Diseases and Conditions, Neurological Disease	28
	Gene Expression, Lipid Metabolism, Small Molecule Biochemistry	21
	Carbohydrate Metabolism, Cardiovascular System	19
	Development and Function, Organismal Development	

Continuation of Table 3.

	Top associated network function	Score
T1-T2	Hematological Disease, Organismal Injury and Abnormalities, Respiratory Disease	44
	Endocrine System Development and Function, Lipid Metabolism, Small Molecule Biochemistry	42
	RNA Post-Transcriptional Modification, Dermatological Diseases and Conditions, Infectious Disease	40
	Cell Morphology, Cellular Assembly and Organization, Cellular Function and Maintenance	33
	Embryonic Development, Nervous System Development and Function, Organ Development	33
T2-T3	Cell Cycle, Cellular Growth and Proliferation, Cellular Development	32
	Lipid Metabolism, Molecular Transport, Small Molecule Biochemistry	27
	Cell Cycle, Cellular Development, Connective Tissue Development and Function	22
	Molecular Transport, Nucleic Acid Metabolism, Small Molecule Biochemistry	20
	Cancer, Carbohydrate Metabolism, Endocrine System Disorders	2

occurring in the loci of significant stage associated aberration were assessed by Chi-square test. The p-values are 0.00025 for all cases and 1.63e-14 for the ER positive subgroup. It suggests that the probabilities finding AD in significant different regions are significantly different among different stages.

Conclusion

To investigate difference for copy number aberration among different stages, we proposed a new approach called GISDIP. To compare well-known GISTIC that identify focal/broad regions of the genome that are significantly amplified/deleted, a

Table 4. IPA outputs of top canonical pathways and top associated network functions for different stages.

Top canonical pathway		-log (B-H p-value)
DCIS-T1 ER+	Retinoic acid Mediated Apoptosis Signaling	1.68
	B Cell Development	1.16
	Autoimmune Thyroid Disease Signalling	1.16
	Allograft Rejection Signaling	1.16
	Graft-versus-Host Disease Signaling	1.16
T1-T2 ER+	Oxidative Phosphorylation	2.37 *
	Notch Signaling	1.89 *
	Natural Killer Cell Signaling	1.88 *
	EIF2 Signaling	1.47 *
	Ubiquinone Biosynthesis	1.19 *
T2-T3 ER+	G-Protein Coupled Receptor Signaling	0.515
	Riboflavin Metabolism	0.515
	Pantothenate and CoA Biosynthesis	0.515
	Propanoate Metabolism	0.515
T2-T3 ER-	Protein Ubiquitination Pathway	1.63
	Aldosterone Signaling in Epithelial Cells	1.03
	NRF2-mediated Oxidative Stress Response	1.03
Top Associated Network Function		Score
DCIS-T1 ER+	Hematological System development and Function, Tissue Morphology, Cellular Growth and Proliferation	45
	Cellular Function and Maintenance, DNA Replication, Recombination, and Repair, Cell Signaling	26
	Molecular Transport, Dematological Disease and Conditions, Developmental	23
	Gene Expression, RNA Post-Transcriptional Modification, Protein Synthesis	41

Continuation of Table 4.

Top canonical pathway		-log (B-H p-value)
T1-T2 ER+	RNA Post-Transcriptional Modification, Dermatological Diseases and conditions, Infectious Disease	39
	Cell Death, Cellular Function and Maintenance, Hematological System Development and Function	32
	Cellular Assembly and Organization, Cellular Function and Maintenance, Dematological Diseases and Conditions	32
	Cellular Development, Reproductive System Development and Function, Connective Tissue Disorders	30
	Carbohydrate Metabolism, Immunological Disease, Lipid Metabolism	30
T2-T3 ER+	Cancer, Connective Tissue Development and Function, Tissue Morphology	25
	Neurological Disease, Developmental Disorder, Embryonic Development	10
	Cancer Reproductive System Disease, Gene Expression	2
	Reproductive System Disease	2
T2-T3 ER-	Reproductive System Disease, Cellular Development, Lipid Metabolism	31

The terms with* indicate -log (p-value) by Fisher's exact test because these were insignificant after applying multiple testing correction.

simulation study could demonstrate each role from GISTIC and GSDIP. GSDIP illustrated different significant regions and biological characters for all samples and for samples stratified

Table 5. Overlapped number of regions obtained by GISDIP on 13K loci obtained by AD in the case of all samples and ER positive samples.

All	#Overlapped (#total)	Chromosome regions detected in 13K loci
DCIS-T1	123 (804)	3q, Xp, q
T1-T2	301 (2938)	1q, 5p, 6p, 8q, 9q, 11q, 13, 14q, 16p, 17q, 19p, 20q
T2-T3	63 (601)	7q, 9p, 11q
ER positive	# overlapped (# total)	Chromosome regions in 13K loci
DCIS-T1	70 (266)	3q
T1-T2	395	1p,q, 2q, 4p, 5p, 6p, 8q, 9q, 11q, 13q, 16p, 17q, 19p,q
T2-T3	60 (261)	--

by ER positive/negative. Significant regions obtained by GISDIP involved the parts overlapped with allele-specific disparity. The probabilities of the overlapping regions were not same between different stages. As the further work, GISDIP could expand to identify differences of expression level like micro-RNA changing according to different stages.

Additional files

[Supplement Table S1a](#)
[Supplement Table S1b](#)
[Supplement Table S1c](#)
[Supplement Table S1d](#)
[Supplement Table S2](#)
[Supplement Table S3](#)
[Supplement Table S4](#)
[Supplement figure S1](#)
[Supplement figure S2](#)
[Supplement figure S3](#)

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Authors' contributions	HKS	HS	FK	DN	ALB	HE	VNK
Research concept and design	✓	✓	--	--	--	--	✓
Collection and/or assembly of data	--	--	--	✓	✓	--	✓
Data analysis and interpretation	✓	✓	✓	✓	--	✓	--
Writing the article	✓	--	✓	--	--	✓	✓
Critical revision of the article	✓	✓	--	--	--	✓	✓
Final approval of article	✓	✓	✓	✓	✓	✓	✓
Statistical analysis	✓	✓	✓	--	--	--	--

Acknowledgement

This study was supported by grants 193387/V50 Understanding breast cancer genomics to ALBD/VKN from the Norwegian Research Council (NFR). We appreciate for Dr. Hege G. Russnes, Department of Genetics, Radiumhospitalet, who gave us valuable biological comments for ER status.

Publication history

Editor: Zhongxue Chen, Indiana University Bloomington, USA.

EIC: Jimmy Efird, East Carolina University, USA.

Received: 20-Sep-2014 Final Revised: 23-Oct-2014

Accepted: 26-Nov-2014 Published: 02-Dec-2014

References

- Beroukhir R, Getz G, Nghiemphu L, Barretina J, Hsueh T, Linhart D, Vivanco I, Lee JC, Huang JH, Alexander S et al., **Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma.** *Proc Natl Acad Sci U S A.* 2007; **104**:20007-12. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Beroukhir R, Mermel CH, Porter D, Wei G, Raychaudhuri S, Donovan J, Barretina J, Boehm JS, Dobson J, Urashima M et al., **The landscape of somatic copy-number alteration across human cancers.** *Nature.* 2010; **463**:899-905. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Riis ML, Zhao X, Kaveh F, Vollan HS, Nesbakken AJ, Solvang HK, Luders T, Bukholm IR and Kristensen VN. **Gene expression profile analysis of t1 and t2 breast cancer reveals different activation pathways.** *ISRN Oncol.* 2013; **2013**:924971. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Solvang HK, Lingjaerde OC, Frigessi A, Borresen-Dale AL and Kristensen VN. **Linear and non-linear dependencies between copy number aberrations and mRNA expression reveal distinct molecular pathways in breast cancer.** *BMC Bioinformatics* 2011; **12**:197. | [Article](#)
- Kaveh F, Edvardsen H, Borresen-Dale AL, V NK and Solvang HK. **Allele-specific disparity in breast cancer.** *BMC Med Genomics.* 2011; **4**:85. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Baumusch LO, Aaroe J, Johansen FE, Hicks J, Sun H, Bruhn L, Gunderson K, Naume B, Kristensen VN, Liestol K, Borresen-Dale AL and Lingjaerde OC. **Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors.** *BMC Genomics.* 2008; **9**:379. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Storey JD. **A direct approach to false discovery rates.** *J.R.Statist.* 2002; **64**:479-498. | [Pdf](#)
- Muggerud AA, Ronneberg JA, Warnberg F, Botling J, Busato F, Jovanovic J, Solvang H, Bukholm I, Borresen-Dale AL, Kristensen VN, Sorlie T and Tost J. **Frequent aberrant DNA methylation of ABCB1, FOXC1, PPP2R2B and PTEN in ductal carcinoma in situ and early invasive breast cancer.** *Breast Cancer Res.* 2010; **12**:R3. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Langerod A, Zhao H, Borgan O, Nesland JM, Bukholm IR, Ik Dahl T, Karesen R, Borresen-Dale AL and Jeffrey SS. **TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer.** *Breast Cancer Res.* 2007; **9**:R30. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Kyndi M, Overgaard M, Nielsen HM, Sorensen FB, Knudsen H and Overgaard J. **High local recurrence risk is not associated with large survival reduction after postmastectomy radiotherapy in high-risk breast cancer: a subgroup analysis of DBCG 82 b&c.** *Radiother Oncol.* 2009; **90**:74-9. | [Article](#) | [PubMed](#)
- Myhre S, Mohammed H, Tramm T, Alsner J, Finak G, Park M, Overgaard J, Borresen-Dale AL, Frigessi A and Sorlie T. **In silico ascription of gene expression differences to tumor and stromal cells in a model to study impact on breast cancer outcome.** *PLoS One.* 2010; **5**:e14002.
- Nielsen HM, Overgaard M, Grau C, Jensen AR and Overgaard J. **Study of failure pattern among high-risk breast cancer patients with or without postmastectomy radiotherapy in addition to adjuvant systemic therapy: long-term results from the Danish Breast Cancer Cooperative Group DBCG 82 b and c randomized studies.** *J Clin Oncol.* 2006; **24**:2268-2275. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Russnes HG, Kuligina E, Susptitsin EN, Voskresenskiy DA, Jordanova ES, Cornelisse CJ, Borresen-Dale AL and Imyanitov EN. **Paired distribution of molecular subtypes in bilateral breast carcinomas.** *Cancer Genet.* 2011; **204**:96-102. | [Article](#) | [PubMed](#)
- Comprehensive molecular portraits of human breast tumours.** *Nature.* 2012; **490**:61-70. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
- Lee EY and Muller WJ. **Oncogenes and tumor suppressor genes.** *Cold*

- Spring Harb Perspect Biol.* 2010; **2**:a003236. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
16. Yokota J. **Tumor progression and metastasis.** *Carcinogenesis.* 2000; **21**:497-503. | [Article](#) | [PubMed](#)
17. Valentin-Vega YA, Barboza JA, Chau GP, El-Naggar AK and Lozano G. **High levels of the p53 inhibitor MDM4 in head and neck squamous carcinomas.** *Hum Pathol.* 2007; **38**:1553-62. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)
18. Hicks J, Krasnitz A, Lakshmi B, Navin NE, Riggs M, Leib E, Esposito D, Alexander J, Troge J, Grubor V, Yoon S, Wigler M, Ye K, Borresen-Dale AL, Naume B, Schlicting E, Norton L, Hagerstrom T, Skoog L, Auer G, Maner S, Lundin P and Zetterberg A. **Novel patterns of genome rearrangement and their association with survival in breast cancer.** *Genome Res.* 2006; **16**:1465-79. | [Article](#) | [PubMed Abstract](#) | [PubMed Full Text](#)

Citation:

Solvang HK, Shimodaira H, Kaveh F, Nebdal D, Børresen-Dale A-L, Edvardsen H and Kristensen VN. **DNA copy number profiles during tumour progression in breast cancer.** *J Med Stat Inform.* 2014; **2**:10.
<http://dx.doi.org/10.7243/2053-7662-2-10>