A telescope for the RNA universe : novel bioinformatic approaches to analyze RNA sequencing data

Pulyakhina, Irina

Chapter 5

Allelic imbalance and its potential role in breast cancer pathogenesis

I. Pulyakhina[1], M.P.G. Vreeswijk[1], J.F.J. Laros[1,2], C.M. Meijers[1], P. Devilee[1], P.A.C. 't Hoen[1]

1 Dept. of Human Genetics, Leiden University Medical Center, Leiden, The Netherlands
2 Leiden Genome Technology Center, Leiden, The Netherlands

Manuscript in preparation

5.1 Abstract

Differential allelic expression (DAE) refers to differences in expression levels between the two alleles of a gene in a cell. DAE can be caused by genetic variation and is known to affect a considerable fraction of genes in the human genome. DAE has been suggested to play an important role in human phenotypic variability, including complex traits and diseases. We hypothesize that DAE contributes to the pathogenesis of breast cancer and that changes in the DAE patterns in control versus breast tumor tissue can be associated with the disease. Deleting or decreasing the expression of the highest expressed allele of a tumor suppressor gene may have larger consequences than deletion of the lower expressed allele.

We analyzed whole transcriptome sequencing data of paired normal-tumor samples from breast cancer patients provided by the TCGA consortium. The allelic expression was examined by analyzing the coverage of the alleles of each variant. We focused on the heterozygous SNPs showing consistent changes in allelic expression in at least 5% of the patient cohort to identify common regulatory mechanisms in breast tumorigenesis.

We introduced a classification of different types of allelic expression changes, and observed that the majority of SNPs showed differential allelic expression only in tumor but not in normal tissue. We found nine SNPs with significant and consistent changes in allelic expression in normal versus tumor RNA. Evaluation of the allelic ratio in tumor DNA demonstrated that the majority of changes in allelic expression was caused by allele-specific aneuploidy. We also found one SNP situated in the *NBPF9* gene that had imbalanced allelic expression already in normal RNA, and this imbalance increased even further in tumor RNA.

The non-random SNP-behavior in the breast cancer patient cohort suggests that DAE is contributing to the mechanism of breast cancer development. DAE and its quantitative effects on gene expression might become the next milestone in understanding breast cancer pathogenesis after the discovery of LOH (loss of heterozygosity) due to aneuploidy at DNA level.

5.2 Introduction

Breast cancer is one of the most important causes of cancer-related deaths, the most common malignancy worldwide in women and the second leading cause of cancer death after lung cancer [198].

Breast tumors are characterized by genomic instability and gross chromosomal aberrations on a genome wide scale, including loss or duplication of whole chromosome arms [199]. Aneuploidy, a process when (parts of) chromosomes can be lost or duplicated, possibly more than once [200, 201], is often observed in breast cancer [202], as well as other types of cancer [203, 204, 205]. The most common consequence of genomic instability and a hallmark of cancer cells is the presence of an abnormal number of chromosomes, or *aneuploidy* [205]. Aneuploidy has been extensively studied and characterized to better understand the cell's evolutionary pathways towards carcinogenesis. Errors in replication, segregation and telomere function are generally assumed to occur in the absence of properly functioning cell cycle checkpoints [206, 207]. During tumorigenesis, DNA regions that include oncogenes are frequently amplified causing their overexpression and endowing the cell a growth advantage; tumor suppressor genes are often lost during the evolution of cancer allowing cells to escape mitotic arrest and/or cell death. Loss of tumor suppressor genes is often a prerequisite for tumorigenesis [208]. Aneuploidy does not affect every chromosome equally; some chromosomes (e.g., chromosome 17 in breast and colon cancer) are more prone to aberrations [209], giving rise to distinct patterns of aneuploidy in certain tumors [210]. An unresolved issue is why there seems to be chromosomal specificity in the pattern of allelic losses and gains in different types of cancer and also in different subtypes of breast cancer [211].

DNA rearrangements can also occur subchromosomally. These rearrangements may result in dis-

torted number of alleles for individual (or sets of neighboring) genes [212]. A frequent phenomenon in breast cancer is loss of heterozygosity (LOH), signifying chromosome regions in the tumor DNA where either the paternal or maternal copy is missing [213, 214, 215]. Simple copy-number loss – monosomy – will therefore show as LOH, but LOH can also be copy-number neutral, when the final number of alleles equals the initial number of alleles [216]. Copy-number neutral LOH happens when one allele is lost and the other allele is gained at the same time [217]. Deviations from the expected 1:1 ratio (different from 1:0 or 0:1) of parental chromosome copies are called allelic imbalance.

Two examples of genes undergoing LOH in sporadic breast tumors are *BRCA1* and *BRCA2*. Regions of the chromosomes where *BRCA1* and *BRCA2* are located (i.e., 17q21 and 13q12, respectively) undergo allelic imbalance or loss of heterozygosity (LOH) in 30-50% of the cases. It was expected that these events would unmask somatic (i.e., acquired) mutations in these genes, but such mutations have rarely been observed [218, 219, 220, 221, 222]. A number of studies have shown that in many individuals the two parental alleles of a gene are not expressed at equal levels [223, 224]. Both *cis*- and *trans*-regulatory mechanisms are at the basis of this phenomenon, which may affect several thousand genes. Genetically, *BRCA1* and *BRCA2* behave like classical tumor suppressor genes, meaning that almost all breast tumors of gene mutation carriers show loss of the wild-type allele [225, 226, 227]. *BRCA1* and *BRCA2* have been implicated in tumorigenesis of sporadic breast cancer as haploinsufficient tumor suppressor genes through the loss of their expression [228, 229]. Mechanistically, this expression loss has been explained either by promoter hypermethylation or LOH, or a combination of both [230]. However, the proportion of tumors with decreased mRNA expression is much higher than the proportion with promotor hypermethylation, while LOH, although found in 30-50% of breast tumors at the *BRCA1* and *BRCA2* loci, reflects a number of different chromosomal mechanisms, not all of which are predicted to lead to expression loss. LOH could affect the expression level of a gene if one allele was physically lost (copy-number loss). If both alleles were expressed at equal levels, this would be predicted to lead to a 50% decrease in expression, but the effect would be stronger if the two alleles of the gene are not expressed at equal levels and the allele with the highest expression is lost. In the latter situation, even copy-number neutral LOH would affect final expression levels.

The examples of *BRCA1* and *BRCA2* discussed above are not allele specific, meaning that either allele can be gained or lost, and such loss leads to reduced and gain to increased gene and protein expression. However, another type of aneuploidy – allele specific aneuploidy, or *allelic disparity* – has also been described [231, 232]. The wild-type allele of the *APC* gene (located on 5q31, associated with hereditary colorectal cancer) was shown to be lost in 22 out of 23 analyzed colon tumors. However, the mutant allele was shown not to contain any typical, inactivating *APC* mutations, but to be expressed at much lower levels than the wild-type allele, leading to lower APC protein expression levels and predisposition to the disease [233]. Unequal expression of the two alleles of a gene, termed *Differential Allelic Expression* (DAE), has also been shown to be associated with breast cancer [234, 235, 236]. Genes as *BRCA1/2, CCND3, EMSY, GPX1, GPX4, MLH3, MTHFR, NBS1, TP53* and *TRXR2* showed distinctive DAE patterns similar among samples of EBV-transformed lymphoblastoid cells [234]. Array-based analysis of eight breast cancer patient-derived normal mammary epithelial lines revealed genes showing DAE to cluster in one major breast cancer-relevant interaction network, which includes two known cancer causative genes, *ZNF331* and *USP6*, and a breast cancer causative gene, *DMBT1* [236].

Thus, aneuploidy plays an essential role in tumorigenesis and breast cancer. Investigating DNA copy number alterations across a tumor's entire genome was a challenging task until the comparative genomic hybridization (CGH) technology was introduced [237]. Subsequently SNP arrays have been used by numerous researchers to explore chromosomal aberrations in tumor DNA, i.e., by comparing a set of healthy control tissues to a set of tumors [238, 239, 240]. With the arrival of Next Generation Sequencing (NGS) and, more specifically, RNA sequencing (RNA-Seq), it became possible to study DAE genome-wide and in genes that were not primary candidates. RNA-Seq is quantitatively more accurate than SNP arrays in determining mRNA levels, so it should be more sensitive to minor changes

in expression. In addition, it facilitates the research on a whole genome level and is not limited to the probe design, as a SNP array is [241]. By comparing genome-wide sequence data (i.e., from normal and tumor RNA of the same patient, or from RNA of healthy subjects and patients with breast tumors), DAE can be investigated for all genes expressed at mRNA level [231, 232, 242, 243, 244].

Some studies have exploited DAE to search for genes in which an allele had been inactivated by protein- truncating mutation, leading to nonsense-mediated mRNA decay. *CHEK2*, a G2 checkpoint kinase 2, showed allelic expression imbalance in 10% of lymphoblastoid cell lines (LCLs) from high-risk breast cancer from whom no mutation in *BRCA1* or *BRCA2* had been identified. All samples with such DAE were carriers of the truncating mutation NM_007194.3:c.1100del [235]. Likewise, in families with pituitary adenoma, a genome-wide scan for genes with DAE in predicted gene carriers led to the discovery of *AIP1* as the culprit underlying gene for this familial cancer syndrome [245]. Individual examples, showing an impact of a certain allelic imbalance event at a patient level, have also been characterized [246]. However, none of the studies so far has attempted to compare DAE of genes with allele-specific patterns of LOH or allelic imbalance in breast tumors, certainly not in a large patient cohort.

In this study, we assessed the phenomenon of allelic imbalance at the DNA level and differential allelic expression at the RNA level. We analyzed a dataset of over 80 breast cancer patients, having data from normal DNA, normal RNA, tumor DNA and tumor RNA from the same patient. The data used in this study was generated within The Cancer Genome Atlas project. We identified coding SNPs in normal DNA to be able to distinguish between the two alleles. We used a combination of statistical tests and developed a statistical framework to assess those SNPs in normal RNA, tumor DNA and tumor RNA from the same patient. We provide a classification of different cases of allele-specific expression using the changes in DAE between normal and tumor RNA. We also address the mechanism causing such DAE patterns using normal and tumor DNA in each patient.

5.3 Materials and methods

5.3.1 TCGA dataset

The analysis was performed on the sequencing data from 89 breast cancer patients having triple-negative breast invasive carcinoma (Figure 5.1, A) available within The Cancer Genome Atlas (TCGA) project. Each patient contributed four samples: normal and tumor DNA samples, and normal and tumor RNA samples. The sample underwent whole exome sequencing. Normal samples were collected from the unaffected part of the patient's breast, while the tumor samples were collected from the breast tumors. DNA/RNA from formalin-fixed, paraffin-embedded tissues were sequenced on the Illumina HiSeq 2000 platform and mapped to the hg19/GRCh37 genome using the BWA aligner [107] for aligning DNA and MapSplice [96] for aligning RNA. The alignment was performed by the TCGA consortium[1]. Level two access data (*bam* files containing aligned reads) were requested from the NCI's Cancer Genomics Hub[2].

5.3.2 Calling and analyzing variants

The workflow of the analysis is represented in Figure 5.1 (panels B and C). Calling variants in patient DNA derived from normal breast tissue has been performed with SAMtools [39] using default parameters. Only heterozygous variants were considered[3]. For that, we selected allelic ratios in normal DNA

[1]http://tcga-data.nci.nih.gov/tcga/

[2]https://cghub.ucsc.edu/

[3]Note that when calling variants in RNA or tumor DNA, heterozygous variants with the unequal coverage of the two alleles are also considered.

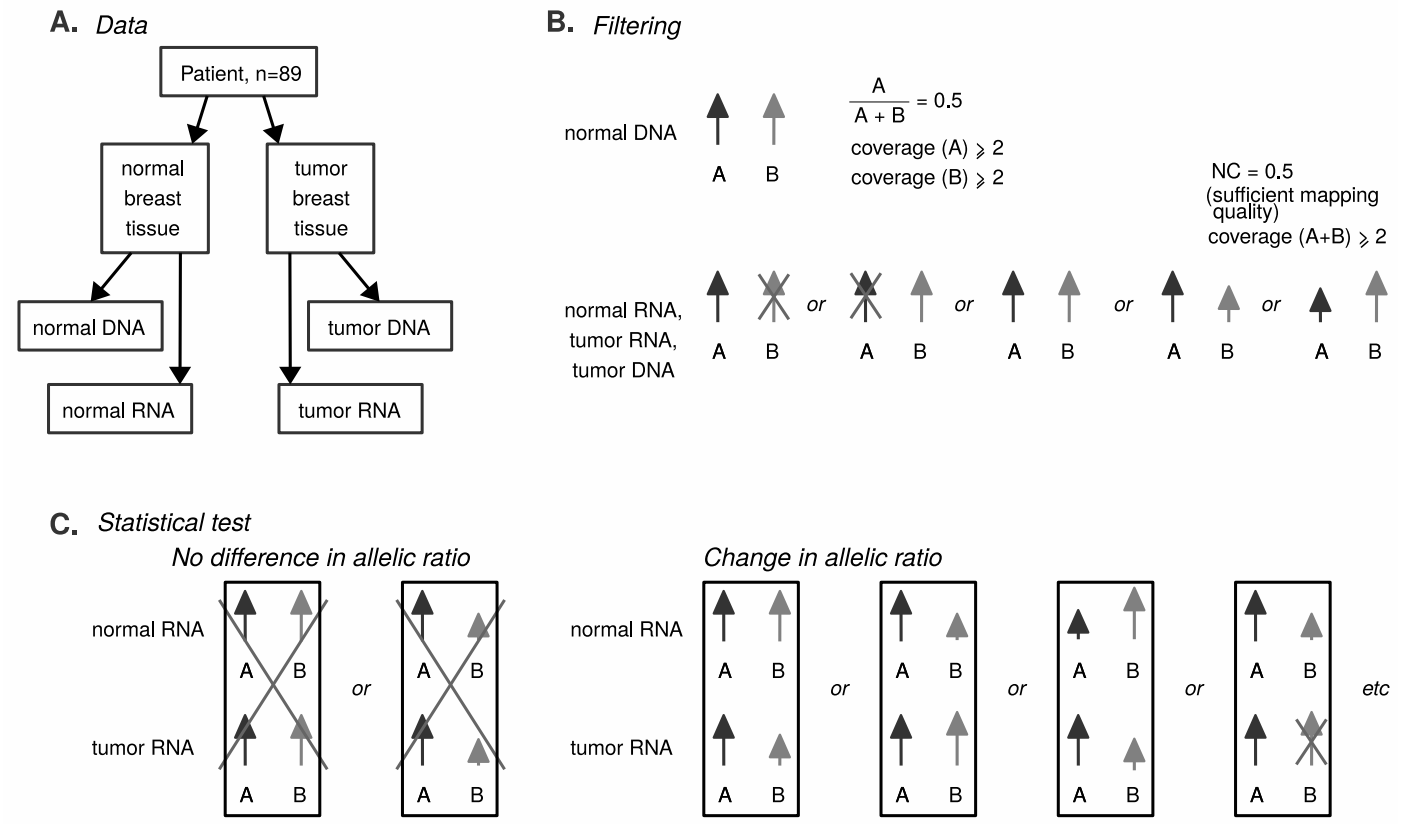

Figure 5.1: **A.** Design of the dataset, in which normal DNA and RNA and tumor DNA and RNA was collected from 89 breast cancer patients. **B.** Filtering steps used during the analysis to select reliable variants. See Section 5.4.2 for more details. **C.** Set-up of the statistical test used to select variants with a change in allelic ratio between normal vs tumor RNA. SNPs showing changes in allelic ratio in normal vs tumor RNA were selected. Note that only recurrent events were selected for the statistical framework to get statistically valid results.

which did not differ significantly from 0.5 (Figure 5.1, B). This was done to exclude variants situated in regions of potential copy number variants (which would result in an allelic ratio different from 0.5), as any change in allelic ratio of these variants in tumors versus normal tissue would be harder to interpret. Each of the alleles also had to be covered at least twice to be selected for further analysis. Calling variants on RNA derived from normal breast tissue has been performed using VarScan [248], reporting homozygous reference, heterozygous and homozygous alternative positions. Only positions where the sum of the coverage of both alleles was above 20 and which had high mapping quality were considered. They were selected using the "NC" flag, variants with "NC=1" (meaning that the position had the mapping quality high enough to be called) were further analyzed. Calling variants on tumor RNA was also performed using VarScan reporting only heterozygous and homozygous alternative variants. The same filtering criteria were used to select high quality variants.

Note that we address variants and SNPs as two separate types of events. We call "a SNP" a nucleotide change on DNA or RNA that has been seen in at least 5% of the cohort, and any other nucleotide change on DNA or RNA is referred to as "a variant".

5.3.3 Statistical framework

We developed a statistical model to identify changes in allelic expression (the ratio between the number of reads mapped to one allele and all reads mapped to this position), which is consistent across multiple samples (Figure 5.1, C). Paired observations – e.g., allelic counts from a normal and a tumor RNA sample – are required. One SNP at a time is tested. The reference allele is referred to as "allele A",

and the alternative allele is referred to as "allele B".

A generalized linear mixed model is fitted for each SNP considering a negative binomial distribution of the allelic counts c. In the model, the fixed effect X is the status of the tissue (normal/tumor) and individuals are modeled as random effects Z:

A generalized linear mixed model is fitted for each SNP considering a negative binomial distribution of the allelic counts c. In the model, the fixed effect X is the status of the tissue (normal/tumor) and the random effect Z is the read counts (different per individual):

$$c = logN + E, \tag{5.1}$$

where c is the coverage of an allele, N is the linear predictor for the response variable and E contains random errors;

$$N = XB + ZU, \tag{5.2}$$

where X is the matrix containing two indicators – showing the tissue state (normal/tumor) and the type of allele (reference/alternative), both can be 0 or 1; Z are the random effects; B and U are the coefficients.

Since the input data is counts, a negative binomial distribution is used as it allows the variance to differ from the mean via introducing an extra parameter of over-dispersion. In the current model, several parameters are estimated, namely the coverage allele A in normal RNA (intercept), the difference in coverage of allele A between tumor and normal RNA, difference between the coverage of allele A and allele B, the interaction between the tissue status (normal/tumor) and the difference between the coverage of allele A and allele B (A 2-way anova model). Four degrees of freedom are needed to estimate these parameters. Two degrees of freedom for the fixed effects, one degree of freedom for the random effect, and one degree of freedom is needed to estimate the standard error. Variants present in at least five patients (20 complete observations: five for each allele in tumor and normal) were selected. The statistical model was implemented in R and is freely available online[4].

To select changes in the expression ratios for non-recurrent events (see Section 5.4.1 for more detail), Fisher's exact test was used (an FDR of 5% was used as a cutoff).

5.4 Results

We assessed variants showing DAE and/or a change in allelic expression in normal versus tumor RNA. At first, we categorized variants (Figure 5.2, A) based on DAE pattern and after that we identified variants present in multiple breast cancer patients and showing a consistent change in RNA allelic expression between normal and tumor tissue. Lastly, we tried to explain the patterns of allelic expression in tumor RNA using tumor DNA information from the same patients.

5.4.1 Variant categories based on DAE

Analyzing 89 breast cancer patients, we identified 78,539 variants heterozygous in at least one patient in normal DNA and present in normal RNA, tumor DNA and tumor RNA (either heterozygous or homozygous). We ran the exact Fisher's test on normal and tumor RNA on each of the variants in each patient individually (which means that the same variant in multiple patients was counted multiple times) and used the FDR cut-off of 5%. This left 5,446 variants present in at least one patient and

[4]https://git.lumc.nl/i.pulyakhina/ase_in_brc/blob/master/TCGA_analysis/

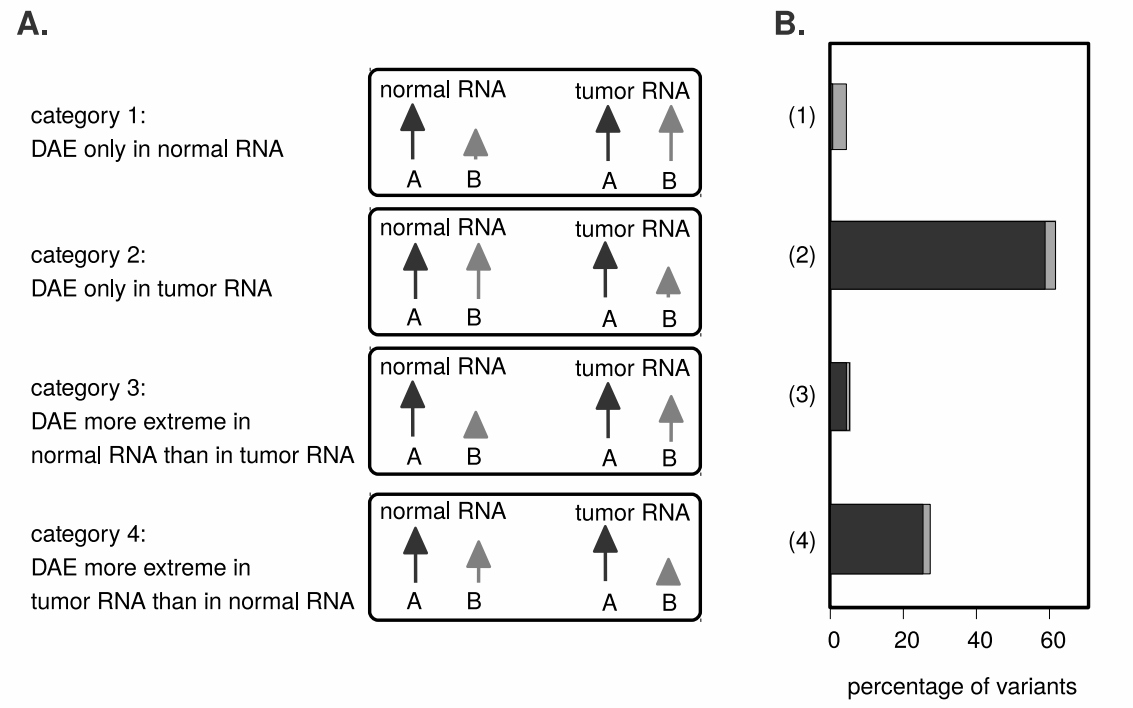

Figure 5.2: Differential DAE in normal or tumor RNA. **A.** Categories of variants based on DAE present in either normal RNA, or tumor RNA, or both. **B.** The percentage of variants present in each category for the variants showing a significant change in the allelic ratio between control and tumor RNA. Note that the fifth category – when both normal and tumor RNA shows DAE to the same extent – is discarded (see Figure 5.1, C) from the analysis and is hence not shown here. Dark grey parts of the bars represent the percentage of variants that showed allelic ratio different from 0.5 in tumor DNA.

showing a change in allelic expression ratio between normal and tumor RNA. We subdivided the variants into the following categories (Figure 5.2, A):

1. DAE only in normal RNA and not in tumor RNA (category 1);
2. DAE only in tumor RNA and not in normal RNA (cat. 2);
3. DAE in both normal and tumor RNA; the imbalance is stronger in normal RNA (cat. 3);
4. DAE in both normal and tumor RNA; the imbalance is stronger in tumor RNA (cat. 4).

Figure 5.2, B depicts the percentage of variants present in each category.

5.4.2 Changes in allelic ratio

Looking at the distribution of the variants across the categories (Figure 5.2), we can appreciate that the vast majority of variants – 89.5% – shows DAE in tumor, i.e., only in tumor – 62% (category 2) – or both in tumor and in normal RNA – 27.5% (category 4). A much smaller percentage of variants shows DAE only in normal RNA (category 1, 4.8% of variants). We can also appreciate that, even when the expression of both alleles is not equal in normal RNA, the imbalance often increases in tumor RNA (category 4) or remains the same (data not shown, as it is not further analyzed, see Figure 5.1, C for details) and rarely decreases (category 3, 5.7% of variants).

We expected aneuploidy to be the major driver of differential allelic expression in tumor RNA, and assessed tumor DNA allelic ratios using two-sided Binomial test against 0.5. When an allelic ratio was significantly different from 0.5 in tumor DNA, we considered that an allele might have been lost or gained. For category 2, 95.7% of the DAE events seen in tumor RNA could be explained by putative aneuploidy in tumor DNA; 97.6% of the variants from category 3 and 97.1% of the variants from category 4. We used category 1 as a control group and expected less allelic imbalance in tumor DNA for the variants from this category, as they did not show any significant imbalance in tumor RNA. As

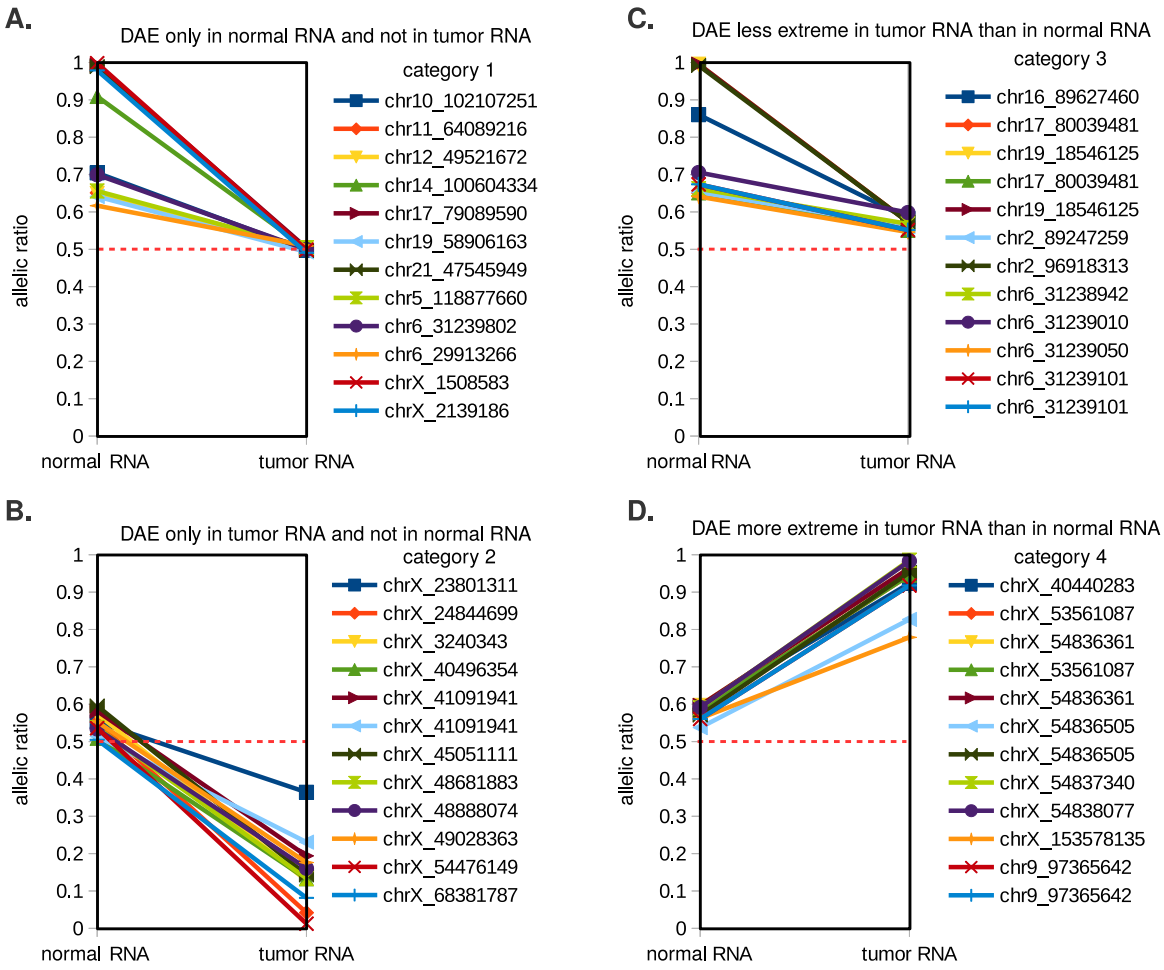

Figure 5.3: Allelic ratio (y axis) and its changes across variant categories introduced in Figure 5.2 in normal vs tumor RNA (x axis). **A.** Variants showing DAE only in normal RNA and not in tumor RNA (category 1 from Figure 5.2). **B.** Variants showing DAE only in tumor RNA and not in normal RNA (category 2 from Figure 5.2). **C.** Variants showing more extreme DAE in normal RNA than in tumor RNA (category 3 from Figure 5.2). **D.** Variants showing more extreme DAE in tumor RNA than in normal RNA (category 4 from Figure 5.2). Each dot represents a variant in one sample, same variant from tumor and normal RNA is connected with a line. Only twelve representative variants are shown per group. Note that we did not distinguish between recurrent and non-recurrent events here. RsIDs are not shown, as not all variants have rsIDs.

expected, only a small amount of variants from category 1 (14.6 %) showed allelic imbalance in tumor DNA (Figure 5.2, B).

We further inspected individual variants present in each category. In Figure 5.3, expression of different alleles (allelic expression ratios) for both normal and tumor RNA are depicted for the different previously introduced classes (Figure 5.2). For Figure 5.3, A, variants show DAE only in normal RNA, which means that the allelic expression ratio can be either <0.5 or >0.5. To make the figure easier to interpret, only variants showing DAE >0.5 in normal RNA are depicted on Figure 5.3. Note that variants showing DAE <0.5 behave similarly to the variants represented in Figure 5.3. For the category of variants where tumor RNA was more imbalanced than normal RNA (category 3 from Figure 5.2; 1,502 variants), the allele ratio was below 0.5 for 627 variants. This indicates that B allele was lost or A allele was gained in tumor compared to normal RNA. For 875 variants the allelic expression ratio was above 0.5 in tumor RNA, which indicates the opposite behavior – the B allele was gained or A allele was lost in tumor compared to normal RNA.

5.4.3 SNPs showing recurrent DAE

After categorizing variants based on DAE, we selected heterozygous SNPs present in multiple patients and consistently showing loss or gain of one of the alleles. We call such SNPs *recurrent*. These SNPs are primary candidates to play a role in the mechanisms of breast tumorigenesis, unlike individual variants which rather explain the cause of the disease in an individual patient [247].

To focus on recurrent SNPs and see how consistent they are across the dataset of 89 breast cancer patients, we selected heterozygous SNPs found in normal DNA of at least five patients (5% of the patient cohort). The statistical framework (described in Section 5.3.3) assigned low p-values to the SNPs that showed significant changes in the allelic expression ratio between tumor and normal RNA, and the change had to be consistent across all patients in which the SNP was expressed. Note that this statistical framework was designed to distinguish between the reference and the alternative allele and to find only cases of allelic disparity. We also performed the analysis to find non-allele-specific aneuploidy events, however, we did not get any significant results different from what is discussed here. From the list of 975 SNPs present in at least 5% of the patient cohort, we detected nine SNPs (Table S1) which showed a recurrent change in allelic expression ratio in tumor versus normal RNA. Figure 5.4 (panel A) contains allelic expression ratios in all patients containing the SNP, and the majority of allelic expression ratios is concentrated around 0.5 in normal RNA and elsewhere in tumor RNA. More information about these nine SNPs can be found in Table S1.

5.4.4 Mechanism behind DAE in tumor RNA

After we identified differences in allelic expression between normal and tumor RNA, we assessed tumor DNA to find an explanation for different allelic expression patterns.

We discovered high correlation between tumor DNA and tumor RNA for six out of nine SNPs (average Pearson's correlation coefficient >0.8). For the same six SNPs, allelic ratio in tumor DNA was significantly different from 0.5 (between 0.67 and 0.98). Both observations indicate the presence of aneuploidy in tumor DNA which is likely to explain the allelic expression observed in tumor RNA. Since allelic expression ratio is calculated as the ratio between the B allele (alternative allele) and the total expression of the variant (Figure 5.4, B), allelic expression ratio equal or higher than 0.5 means that the alternative B allele was gained (or the reference A allele was lost) in all patients. The events displayed in Figure 5.5 are examples of allele-specific aneuploidy, previously introduced as allelic disparity – when one specific allele is consistently gained or lost in all the patients. We observed that the same allele that was lower expressed in normal RNA was gained in tumor DNA and became higher expressed in tumor RNA (Figure 5.5, A, three panels show the data for an example SNP, rs946837, *TMTC4*). We also observed one SNP which showed the opposite pattern – the allele that was higher expressed in normal RNA was lost in tumor DNA and became lowew expressed allele in tumor RNA (data not shown, as it resembles Figure 5.5, A, but with a decreasing slope).

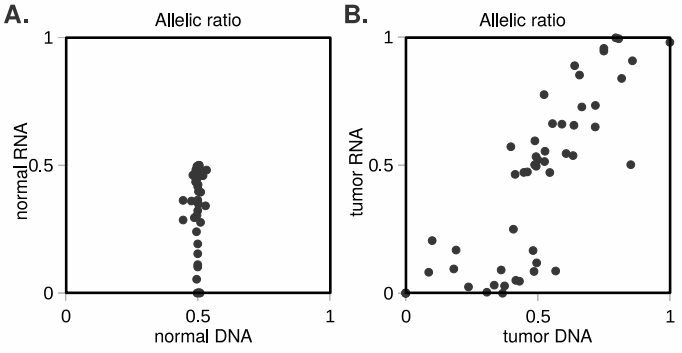

Figure 5.4: **A.** Scatter plot showing the distribution of allelic ratios for the nine SNPs showing significant changes of allelic ratio in normal vs tumor RNA. x axis shows the allelic ratio in normal DNA. y axis show the allelic ratio in normal RNA. **B.** Correlation between the allelic ratio in tumor DNA (x axis) and tumor RNA (y axis). Allelic ratio on plots A and B is calculated as the expression of the alternative allele divided by the total coverage of the allele. Each dot represents one sample, meaning that the same SNP coming from different patients is present multiple times. See Figure 5.5 for the examples of individual SNPs.

We observed a different behavior of allelic expression for the other three SNPs that showed different allelic expression in tumor RNA versus normal RNA (Figure 5.5, B, three panels show the data for an example SNP, rs2798893, *NBPF9*). All three SNPs had coverage above nine in all samples, and for the vast majority of patients the coverage in each sample was above 20. However, at the RNA level we could see that the allele that was lower expressed in normal RNA lost even more expression in tumor RNA (and the allelic expression ratio went down) (for rs7185949, *USP10* and rs2798893, *NBPF9*). A slight (but consistent) increase of the lower expressed allele was observed for rs617073 situated in the *U2AF2* gene.

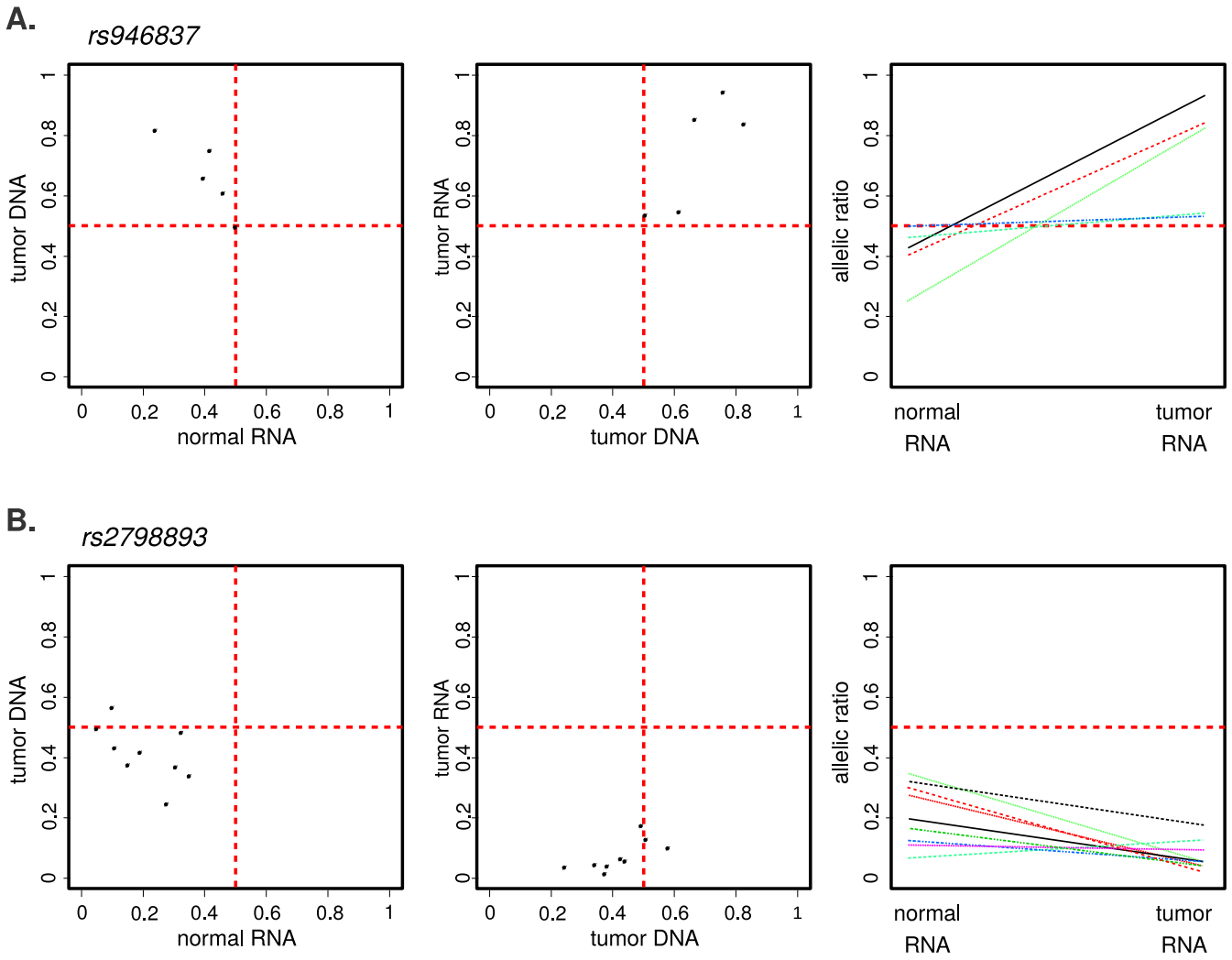

Figure 5.5: **A.** Example SNP representing SNPs with potential allelic disparity. **B.** Example SNP representing the rest of identified SNPs. Left panel – correlation between the allelic ratio in normal RNA (x axis) and tumor DNA (y axis). Middle panel – correlation between the allelic ratio in tumor DNA (x axis) and tumor RNA (y axis). **C.** Changes in the B allele frequency in each pair of samples in normal vs tumor RNA. Each dot represents a variant in one sample, same variant from tumor and normal RNA is connected with a line.

5.4.5 Allelic imbalance in normal RNA and its behavior in tumor RNA

The previous paragraph highlighted SNPs (present in at least five individuals) that showed consistent changes in allelic ratio of RNA in tumor versus normal tissue state. The majority of variants that we identified showed no DAE in normal RNA, meaning that both alleles were equally expressed. Here, we focused on the variants present in at least 5% of the samples, that already have DAE in normal

RNA and show a change in the allelic expression ratio in tumor RNA. We found almost 50% of variants (12,011 of 29,140 variants) imbalanced in normal RNA to be detected only in one sample (Figure 5.5, A and B). Only 15% of the variants (4,348 variants) were found in at least 5 samples with DAE in normal RNA. We assessed whether any recurrent change in the allelic ratio between normal and tumor RNA within these 4,348 variants were found.

Out of these 4,349 SNPs, we discovered only one SNP – already mentioned above, rs2798893 in the *NBPF9* gene – imbalanced in normal RNA to be present in at least 5% of breast cancer patients, which showed a significant change in DAE in tumor vs normal RNA. This SNP was also identified previously without pre-selecting SNPs with DAE in normal RNA before running the statistical framework. Analyzing tumor DNA and RNA profiles, we did not see any correlation between the allelic imbalance in tumor DNA and allelic expression ratio in tumor RNA ($\rho = 0.45$). We observed allelic expression patterns depicted on Figure 5.5 (panel B) – frequencies of the lower expressed allele in tumor DNA were mainly concentrated in the window of [0.4, 0.6]. The majority of nine samples heterozygous at this SNP showed a consistent decrease of the expression of the allele that was lower expressed in normal RNA.

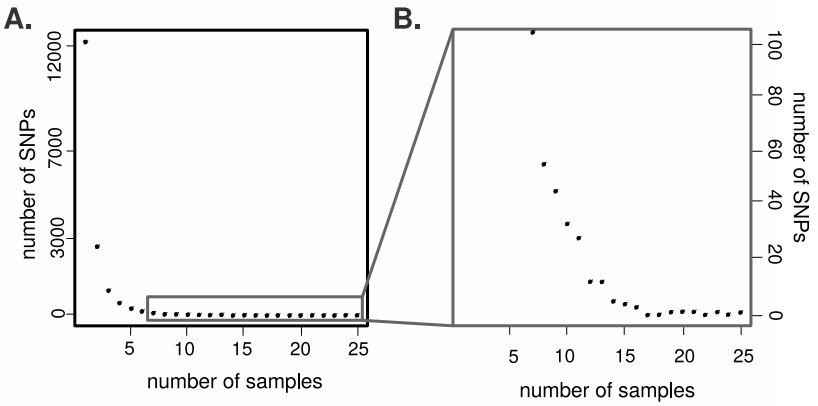

Figure 5.6: Distribution of variants showing DAE in normal RNA across the 89 samples of breast cancer patients. **A.** Number of samples (x axis, i.e., "1" on the x axis means that the variant has to been seen only once and not more) containing a certain number of variants (y axis). **B.** Zoomed-in version of A, y axis is restricted by 100 variants.

5.5 Discussion

Aneuploidy, LOH and their role in breast cancer have extensively been studied over the past decades. Resulting in the unequal expression of the two alleles, aneuploidy was shown to lead to reduced gene expression [230]. Genome-wide research of aneuploidy in breast cancer was conducted mainly using CGH arrays and exome- sequencing with Next Generation Sequencing (NGS) technology and resulted in numerous regions showing genomic instability in multiple or individual breast cancer patients. Earlier considered not to be allele-specific, meaning that either of the two alleles can be gained (or lost) with the same probability, aneuploidy was recently shown to occur in allele-specific manner [220, 242, 243]. One would expect that loss or gain of an allele in tumor DNA would have a direct influence on the gene expression in tumor RNA, which indeed has been shown [220]. However, unequal RNA expression of alleles, or Differential Allelic Expression (DAE), present already in normal RNA – and its changes in breast cancer tumor RNA – has not been extensively studied yet.

In this study we performed a thorough examination of a breast cancer patient cohort provided by The Cancer Genome Atlas project. Leaving the somatic mutations behind, we explored SNPs present in normal DNA and RNA of each patient individually and examined the behavior of these SNPs in the tumor DNA and RNA of the same patient. We observed that the majority (over 50%, data not shown) of variants occurred in normal and tumor DNA and RNA of only one patient and showed equal expression of both alleles in normal RNA and DAE in tumor RNA. We also found variants (33.2%) that showed DAE in both normal and tumor RNA, 27.5% of detected variants showed a stronger allelic imbalance (further away from 0.5) in tumor RNA and only 5.7% of detected variants showed a stronger imbalance in normal RNA.

We assumed that the majority of the SNPs which showed DAE in tumor RNA can partly be explained by aneuploidy, which we tested using tumor DNA of the same patients. We could indeed see that over 95% of variants which showed DAE in tumor RNA also showed allelic instability (allelic ratio different from 0.5) in tumor DNA.

Despite a lot of variation between the patients and the presence of patient-specific variants, we were able to identify 975 heterozygous SNPs present in more than 5% of the patients. We discovered that nine of these SNPs showed a significant and consistent change in allelic expression between normal and tumor RNA. For six out of these nine SNPs, showing allelic expression ratio significantly different from 0.5 in the majority of patients and high correlation between tumor DNA and tumor RNA (average Pearson's correlation coefficient >0.8), suggested that aneuploidy could be one of the potential contributors to DAE in tumor RNA. Moreover, from the values of allelic ratios in tumor DNA and allelic expression in tumor RNA of these SNPs, we derive that one specific allele was gained or lost, which suggests that these SNPs are probable cases of allele-specific aneuploidy (or allelic disparity). Such behavior – the presence of DAE and the absence of aneuploidy – might for instance be explained by allele-specific binding of transcription factor proteins (that would explain DAE in normal RNA), for which the expression is altered in breast tumors (that would explain changes in DAE in tumor versus normal RNA). Further investigation is necessary in order to explain the mechanism driving allelic disparity.

Discovering a very limited number of recurrent events is an expected outcome, as is has already been shown for a number of different cancer types including breast cancer that, i.e., for somatic mutations only a few altered genes are recurrent in at least 10% of the patients, while the vast majority of changes happens in less than 5% of the samples [249, 250]. Another study which focused on the analysis of DAE in a limited set of just eight breast cancer patients was only able to find three genes – *ZNF331, USP6* and *DMBT1* to show a change in DAE in tumor vs normal cDNA. Our analysis assessed individual SNPs (which we used as the markers for different alleles), as we hypothesized that specific gene alterations recurrent in the patient cohort might have a contribution to breast cancer pathogenesis. As the number of identified recurrent SNPs is limited, we suggest that we should rather look to the patterns observed for larger chromosomal regions, as certain parts of chromosomes are expected to show similar aneuploidy pattern. However, a recent study [251] suggests that, as recurrent events are rare due to the nature of cancer, a more meaningful analysis should be conducted on a level higher than gene or locus, which can be done via phasing the genotypes. A pathway or network analysis can help us to effectively uncover biological systems perturbed in tumor cells which is unfortunately still a very challenging analysis and has to be developed and studied more extensively.

Our understanding of aneuploidy and its contribution to breast cancer pathogenesis, as well as LOH and DAE, is still not full despite the decades of analyses on various levels. However, the increasing power of bioinformatic analysis transcriptome- and genome-wide and a potential of extensive research on large patient cohorts make a great promise for understanding and explaining breast cancer development and pathogenesis.

5.6 Acknowledgements

The results published here are in whole or part based upon data generated by The Cancer Genome Atlas pilot project established by the NCI and NHGRI. Information about TCGA and the investigators and institutions who constitute the TCGA research network can be found at http://cancergenome.nih.gov.

We would like to thank E.J. de Meijer and H.P.J. Buermans for their valuable advice and useful discussions, A. Morra and L. Eijsackers for the development of the statistical framework, M. van Iterson for his help with using the BiocParallel R package and M.P. Villerius for his help with running the analysis on the cluster. We also would like to acknowledge the TCGA consortium for providing us with the data of the breast cancer cohort.

5.7 Appendix

Table S1: Description of nine identified SNPs described in the main text.

Chrom.	Position	rsID	Gene name	ρ	Gene description
chr1	144,931,392	rs2798893	NBPF9	0.45	neuroblastoma breakpoint family member 9, higher expressed in breast, ovarian and pancreatic cancer
chr1	201,112,981	rs8158	TMEM9	0.84	a positive regulator of neurite outgrowth and its overexpression was suggested to play primary role in prostate carcinogenesis
chr9	96,238,578	rs10821135	FAM120A	0.98	a constitutive coactivator of PPAR-gamma-like protein 1, which has a higher expression in numerous cancers (including breast cancer)
chr12	9,232,268	rs669	A2M	0.72	alpha-2-macroglobulin which was suggested to play a role in inhibiting brain tumors
chr13	101,287,340	rs946837	TMTC4	0.67	transmembrane and tetratricopeptide repeat containing 4 known to be expressed in breast (mammary gland) tumor
chr16	68,857,441	rs1801552	CDH1	0.95	cadherin-1 protein, for which the increased risk of having breast cancer has been shown in case of genetic variations
chr16	711,905	rs2301426	WDR90	0.69	WD repeat-containing protein 90, higher expressed in about 45% of known cancer types (excluding breast cancer)
chr16	84,779,248	rs7185949	USP10	0.21	tumor-associated marker in gastric carcinoma
chr19	56,180,968	rs617073	U2AF2	0.96	U2 small nuclear RNA auxiliary factor 2, has not been associated with any type of cancer before