Carcinogenesis Advance Access published August 29, 2015

Genomic copy number analysis of Chernobyl papillary thyroid carcinoma in the Ukrainian-American Cohort

Selmansberger M[1], Braselmann H[1], Hess J[1], Bogdanova T[2], Abend M[3], Tronko M[2], Brenner A[4], Zitzelsberger H[1], and Unger K[*1]

[1]Research Unit Radiation Cytogenetics, Helmholtz Zentrum München, German Research Center for Environmental Health GmbH, Neuherberg, Germany
[2]Institute of Endocrinology and Metabolism, National Academy of Medical Sciences of the Ukraine, Kiev, Ukraine
[3]Bundeswehr Institute of Radiobiology, Munich, Germany
[4]Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, NCI, Bethesda, MD, USA

July 6, 2015

Abstract

One of the major consequences of the 1986 Chernobyl reactor accident was a dramatic increase in papillary thyroid carcinoma (PTC) incidence, predominantly in patients exposed to the radioiodine fallout at young age. The present study is the first on genomic copy number alterations (CNAs) of PTCs of the Ukrainian American cohort (UkrAm) generated by Array Comparative Genomic Hybridization (aCGH). Unsupervised hierarchical clustering of CNA profiles revealed a significant enrichment of a subgroup of patients with female gender, long latency (> 17 years) and negative lymph node status. Further, we identified single CNAs that were significantly associated with latency, gender, radiation dose and BRAF status. Multivariate analysis revealed no interactions but additive effects of parameters gender, latency and dose on CNAs. The previously identified radiation-associated gain of the chromosomal bands 7q11.22-11.23 was present in 29% of cases. Moreover, comparison of our radiation-associated papillary thyroid carcinoma (PTC) data set with the TCGA data set on sporadic PTCs revealed altered copy numbers of the tumor driver genes NF2 and CHEK2. Further, we integrated the CNA data with transcriptomic data that were available on a subset of the herein analyzed cohort and did not find statistically significant associations between the two molecular layers. However, applying hierarchical clustering on a "BRAF-like/RAS-like" transcriptome signature split the cases into four groups, one of which containing all BRAF-positive cases vali-

*unger@helmholtz-muenchen.de

dating the signature in an independent data set.

1 Summary

This study reports on genomic copy number alterations (aCGH) in the post-Chernobyl Ukrainian-American PTC cohort and a comparison of the results with previously published knowledge on post-Chernobyl PTCs and with a comprehensive multi-level omics study on sporadic PTCs.

2 Introduction

One of the major consequences of the 1986 Chernobyl nuclear accident to human health was a dramatic increase in thyroid cancer incidence among those who were children or adolescents at the time of exposure [1, 2]. Several studies clearly related this increase to radioiodine exposure (mainly I-131) from the reactor accident [3, 4, 5, 6, 7]. Numerous genomic studies on post-Chernobyl papillary thyroid carcinomas (PTCs) have been published so far. A study by Unger et al. (2004) has shown intratumoral heterogeneity of RET/PTC rearrangements in post-Chernobyl PTCs suggesting either a multiclonal origin of the tumors or RET/PTC being a late subclonal event [8]. A comparative study on global genomic copy number alterations (CNAs) in two age-matched cohorts of post-Chernobyl PTCs and sporadic PTCs revealed a radiation-specific genomic copy number gain of the chromosomal bands 7q11.22-11.23, which was exclusively detected in exposed cases by array comparative genomic hybridization (aCGH) [9]. Further, a significant mRNA overexpression of the CLIP2 gene, located in the gained band, was shown [9]. Amongst others, the largest cohort of subjects who were exposed to the radioiodine-contaminated fallout from the Chernobyl accident under the age of 18 is the so called Ukrainian-American (UkrAm) cohort [10] which was established by the US National Cancer Institute and the Ukrainian National Academy of Medical Sciences [11]. All subjects of this cohort received direct thyroid I-131 measurements within two months after the accident, which built the basis for calculating individual dose estimates. Along with comprehensive epidemiological data available UkrAm represents an excellent cohort for the investigation of radiation-associated thyroid cancer risk. As a result of four screening examinations conducted between 1998 and 2007 on approx. 13,000 subjects a total number of 110 thyroid cancer cases were detected, the majority of which were PTCs [11]. Dose estimates for the entire UkrAm cohort showed an arithmetic mean of 0.68 Gy, a geometric mean of 0.23 Gy, and a lognormal distribution [12, 13]. At present, there are two transcriptomic studies on the UkrAm cohort published by Abend et. al (2012) [14, 15] describing dose-dependent mRNA expressions in normal tissue and tumor tissue, respectively. However, so far no comprehensive studies on the genomic level have been conducted on the UkrAm cohort. Hence, the present study on global genomic copy number changes in PTCs from patients of the UkrAm cohort complements the existing studies. The main aim of this study was an explorative investigation of CNAs in 84 PTCs of the UkrAm cohort and of possible associations with clinical parameters and patient data including individual dose estimates. The results provide the basic knowledge required for an integration of molecular data on radiation-associated PTCs into epidemiology-based modeling of radiation-associated papillary thyroid cancer risk.

3 Material and Methods

3.1 Patient and tissue samples

We analyzed the tumor samples from 84 PTC patients of the Ukrainian-American cohort (UkrAm). The tissue samples were collected by the Chernobyl Tissue Bank (CTB) after surgical removal of the thyroid gland due to PTC diagnosis [11]. The CTB pathology panel examined all PTC tissues that were provided by the CTB. The entire UkrAm cohort comprises 13,243 persons from the Ukraine who were younger than 18 years at the time of the Chernobyl nuclear accident in April 1986. From all members, an individual direct thyroid dose measurement was conducted within two month after the accident. Individual doses were estimated and published by Likhtarov et al. 2006 [12]. Between 1998 and 2008, 110 thyroid cancers were diagnosed, including 104 PTCs.

3.2 DNA and RNA samples

Prior to nucleic acid extraction all FFPE tissue sections were macro-dissected in order to maximize the proportion of tumor cells in the analyzed tissues. The resulting cellularities i.e. proportion of tumor cells per tissue ranged between 50 and 90%. All DNA and RNA samples analyzed in this study were extracted from formalin fixed paraffin embedded (FFPE) tissue sections, using the Qiagen AllPrep DNA/RNA FFPE kit (Venlo, Netherlands) according to the manufacturer's protocol. Quantification of the nucleic acid samples was conducted with a NanoDrop spectrophotometer (Thermo scientific, Waltham, USA).

3.3 BRAF V600E mutation and RET/PTC rearrangements

Data on the BRAF V600E mutation status and RET/PTC rearrangements were included as published by Selmansberger et al. (2014) [16]. Additional data summarized in Leeman-Neill et al. (2013) [17] were added for individual cases after personal communication with Prof. Yuri Nikiforov (Division of Molecular Genomic Pathology, University of Pittsburgh, U.S.).

3.4 Array CGH analysis and hierarchical clustering

Array CGH profiles were generated using Agilent Sure Print 60K oligo microarrays (Agilent, Santa Clara, CA, USA, AMADID 021924). DNA labeling and purification was carried out as described by Hess et al. (2011) [9]. Quality of labeled DNA was assessed by determining the "specific labeling efficiency", which had to exceed incorporation of a minimum of 15 pmol dye per μg DNA and the labeling yield which had to be greater than 1 μg. A detailed description of the QA measures applied is given by Buffart et al. (2007) [18]. Hybridization and washing of arrays and extraction of raw data was performed according to the manufacturer's protocol. Raw data quality assessment, data preprocessing, and subsequent generation of copy number profiles were performed using the statistical platform R and the MANOR, CGHbase, CGHcall (provided with cellularities from histology data), and CGHregions packages, available from Bioconductor [19, 20, 21, 22]. A detailed description of data preprocessing and data analysis can be found in [23]. Unsupervised hierarchical clustering of the obtained aCGH profiles was performed on copy number calls using Euclidean dis-

tance and was visualized using an in-house written function.

3.5 Gene expression data

Genome wide gene expression profiling data were generated using Agilent 44K oligo microarrays by the Bundeswehr Institute of Radiobiology and raw data were provided in personal communication (Abend et al., 2012). Details are described by Abend et al. 2012. Quality assessment and preprocessing of the mRNA expression data was carried out using the statistical platform R and the limma package and the Agi4x44PreProcess package available from Bioconductor (Lopez-Romero; Smyth, 2005). A total number of 31 UkrAm cases was included in the data set, while 23 of which also were present in our aCGH data set which allowed systematic integration of the two data layers. The gene expression data set was subjected to unsupervised hierarchical clustering using 59 genes out of the 71 "BRAF-like"-signature genes as published in a large study (n=391) on sporadic PTCs (Cancer Genome Atlas Research, 2014).

3.6 Association of genomic copy number alterations with clinical data

Univariate testing

The initial association testing was performed using the CGHtest approach (`http://www.few.vu.nl/~mavdwiel/CGHtest.html` [24]), that is based on permutation t-testing. We conducted our calculations with 10,000 permutations. All available clinical data (i.e. age at exposure, age at operation, latency, histological subtype, TNM status, thyroid dose) were tested for associations with CNAs. For continuous variables such as age or thyroid dose, fixed thresholds were used for dichotomization and the resulting groups were subsequently tested for significant differences in the frequency of copy number changes for each region. The thresholds for the parameter latency was set to 17 years, which is in accordance with the median (16.6 yrs) and mean (17.4 yrs) latency of the investigated samples. Thresholds of 5 years for age at exposure (AaE) and 20 years for age at operation (AaO) were taken from Selmansberger et al. [25]. For the association with thyroid dose, three groups were formed according to Abend et al. [15] i.e a low-dose group (< 300 mGy), an intermediate dose group ($300 - 1000$ mGy), and a high dose group (> 1000 mGy).

Multivariate testing

A substantial overlap of significantly associated regions between the parameters dose, gender, and latency was the motivation to assess the statistical interactions between these variables. The multivariate analysis was conducted using a logistic two-way model for the binary variable Y (gain vs. no-gain / loss vs. no-loss). For this purpose the model $Y \sim A + B + A * B$ (model 1) was tested, were Y is the dependent variable (i.e. occurrence of the CNA), A and B are the two parameters, and A*B is the interaction of the two parameters. Significance of the A*B interaction term was tested by the comparison to the pure additive model $Y \sim A + B$ (model 2) to the full model 1 (with interaction term) with respect to their deviance difference [26]. First, for each CNA region the significance of the full model was tested and the p-values were FDR adjusted using the Benjamini-Hochberg approach [27]. In a second step, the significance of the interaction term of model 1 was tested for those CNA regions with a significantly better fit of model 1 compared to the null model. If there was no significant inter-

action observed, the additive contributions of A and/or B were assessed with model 2. P-values were adjusted for FDR and significance was generally accepted for FDR ≤ 0.05. Details on the multivariate testing approach can be found in SI addendum 1.

3.7 Dosimetry

Individual thyroid doses on the investigated subset of 84 cases were used as published previously [12, 28, 13]. The doses are summarized in Table 1 and shown as categorized values for individual cases in SI Table 1 and Figure 2.

3.8 Unsupervised hierarchical clustering based on mRNA expression data and comparison with TCGA study on sporadic PTC

The TCGA study presented a so-called "BRAF-RAS score", enabling the classification of PTC cases into "RAS-like" or "BRAF-like" [29]. A gene signature composed of 71 genes that built the basis for calculating the BRAF-RAS score was derived from a differential expression analysis between a set of tumors with BRAF V600E mutation and a set of tumors with RAS mutation. 59 out of these 71 genes were also present in our available gene expression microarray data set and we subjected the appropriate expression values of this subset to unsupervised hierarchical clustering. Subsequently, we tested possible associations of the parameters, gender, lymph node status, latency, age at operation (AaO), age at exposure (AaE), CLIP2 marker status, BRAF V600E mutation status, RET/PTC1 and RET/PTC3 status and dose with grouping of cases according to clusters of the first and second hierarchy (see Figure SI 1). For this, Fisher's exact test was used.

3.9 Comparison of CNAs in UkrAm PTCs and sporadic PTCs of TCGA cohort

In order to compare the genomic copy number results of our study with the data on sporadic PTCs published by the TCGA consortium [30] we downloaded the latest analysis version (04/02/2015) of the GISTIC2 typed copy number data of the TCGA papillary thyroid cancer data set ("THCA") from the Broad Institute server using the "firehose get" command line binary. The sample barcodes of the data that were included in the [30] paper were downloaded from CBioportal [31] as part of the clinical data table. In order to match the TCGA data set to our data set, only data from patients younger than 35 years at the time of diagnosis were included, resulting in a subset of 113 patients. Matching according to histological data was not possible since different histopathology classifications were used. The clinical data table for the TCGA data subset can be found in the supplementary Table 5. The GISTIC2 called copy number data retrieved via "firehose" were subjected to the following comparisons with our data set: a) Overlap between regions statistically significantly associated with BRAF mutation status (our data) or "BRAF-like" status (TCGA data), b) overlap between regions statistically significantly associated with gender in both data sets c) presence of gain of CLIP2 or the 7q11.22-11.23 region in the TCGA subset d) presence of the groups "isolated loss of 22q", "gain of 16q" and "silent SCNA" according to the classification given in [30] in the UkrAm copy number data set. Association testing of copy number alterations was performed using the R package CGHtest (`http://www.few.`

vu.nl/~mavdwiel/CGHtest.html) whilst associations were considered as statistically significant if both the p-value and the false-discovery rate were below 0.05. Frequencies of alterations of the TCGA THCA subset were plotted using a customized version of the function "frequencyPlot" from the CGHregions package [21].

3.10 Integrative analysis of aCGH profiles and global mRNA expression data

The integration of 24 aCGH profiles and the corresponding global mRNA expression data was carried out with tailor-made approaches implemented in the sigaR R package [32, 33] using the default values provided by the package for all parameters during the different analysis steps.

4 Results

4.1 BRAF V600E mutation and RET/PTC rearrangement rates

From the complete data set, a total number of 70 cases were typed for RET/PTC1 and RET/PTC3 rearrangement status. Overall, 31% of the cases showed a rearranged RET gene whilst 18 out of 70 cases showed RET/PTC1- and 4 out of 70 cases showed RET/PTC3 rearrangement. From 70 cases we obtained valid BRAF V600E Sanger sequencing profiles, whilst 9 out of 70 cases showed a BRAF V600E mutation (13%).

4.2 aCGH analysis

High-resolution aCGH profiles from 84 PTC samples were generated. All 84 cases fulfilled the quality criteria after isothermal amplification. Figure 1 shows the cumulative frequencies of CNAs in the complete data set. A subset of 24 out of 84 (29%) cases showed a copy number gain of the chromosomal bands 7q11.22-11.23. The most frequently detected DNA gains were on chromosome 11q with 44% and chromosome 19q with 42% of cases harboring the gain (see SI Table 2: region No. 258/259, region No. 405). The most frequently detected DNA loss was on chromosome 16q with 33% of cases showing the loss. Detailed information on all detected CNAs are listed in SI Table 2 and SI Table 3.

4.3 Unsupervised hierarchical clustering of aCGH profiles

Figure 2 shows the result of the unsupervised hierarchical clustering of 84 aCGH profiles using Euclidean distances, the distribution of the sample parameters gender, lymph node status, latency (categorized), AaO, AaE, CLIP2 marker status, BRAF V600E mutation status, RET/PTC status, and dose. The CLIP2 marker status was assessed as described in detail by Selmansberger et al. (2014) [16]. The clustering revealed two main clusters (C1 and C2) and three parameters were significantly differentially distributed across C1 and C2. Firstly, there were more females in C1 compared to C2 ($p=0.0054$, Exact Fisher's Test), secondly, more patients with lymph node metastases were in C2 compared to C1 ($p=0.025$), and thirdly more patients with latency greater than 17 years in C1 compared to C2 ($p=0.0028$). Interestingly, the clustering can be completely rebuilt by only considering changes of chromosomes 1p, 19. Thus, profiles of cluster C1.1 showed no copy number change of chromosome 1p but gain of chromosome 20. For cluster C1.2 all but one profile

showed gains of both 1q and chromosome 20. Profiles of cluster 2.1 do not show any of the mentioned alterations, whilst cluster 2.2 only shows a loss of chromosome 19. These two regions, therefore, seem to be most informative for the grouping according to unsupervised hierarchical cluster analysis.

4.4 Association of copy number alterations with patient data

Significant testing results from the univariate approach ($p < 0.05$ and FDR < 0.05) were obtained for the parameters BRAFV600E mutation status, latency, gender, and dose. With regard to BRAF V600E mutation gain of 6 regions on chromosomes 1 and 5 were associated with positive BRAF V600E status. Although the result was statistically significant, the finding is based on only 9 samples with a mutation of BRAF V600E and was therefore not suitable for multivariate analysis. An overview on the univariate testing results of parameters latency, gender and dose can be found in Table 2. 52 gained regions on chromosomes 1, 2, 3, 4, 6, 7,10, 11, 12, 15, 16, 17 and 19 and loss of a region on chromosome 7p were associated with female gender. 161 gained and 26 deleted chromosomal regions on chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 14, 15, 16, 17, 19, 20, 21, and 22 were associated with long latency and 128 gained regions on chromosomes 1, 2, 3, 4, 5, 6, 7, 8, 9, 11, 12, 14, 15, 16, 17, 19, 20, 21 and 22 were associated with the intermediate dose group. We observed many chromosomal regions commonly associated with BRAF V600E status, dose, gender and latency which motivated us to analyze these parameters with a multivariate approach in order to test for possible statistical interaction or additive effects. Statistical interaction describes the simultaneous influence of two or more parameters on a third variable (i.e. the occurrence of CNAs) that is non-additive. The multivariate approach revealed no significant interactions between CNA associated parameters latency, gender, and dose. However, the majority of copy number alterations were determined by additive effects of the parameters dose, gender and latency (see Suppl. Table 2). In detail, out of the total number of 215 CNAs that were associated with BRAF V600E, gender, latency or dose, 139 copy number regions were not determined by additive effects of any of the parameters, 5 by simultaneous additive effects of dose/gender, latency/gender and latency/dose, 44 by dose/gender and latency and gender, 13 by latency/dose alone and 9 by latency/gender alone. With regard to copy number losses, 205 regions were not determined by additive effects at all and 10 regions by latency/dose alone. For a detailed overview of all statistical significant regions from the univariate or multivariate analysis see SI Tables 2 - 6.

4.5 Comparisons with findings from the TCGA study on sporadic PTCs

We compared our findings on genomic copy number alterations with that published on a large study on sporadic PTCs conducted by the TCGA consortium (Cancer Genome Atlas Research, 2014). Including only patients younger than 35 years at the time of diagnosis resulted in a subset of 113 cases. The cumulative frequencies of all CNAs in this subset are shown in Figure SI 2. The majority of cases (94 out of 113) showed no CNAs at all and correspond to the "SCNA-quiet" category which also was present in 45 out of the 84 UkrAm cases. 16/113 cases (14%) showed a copy number loss on chromosome 22q

only ("SCNA-22q-del") and 3/113 cases (2.7%) a copy number gain on 1q ("SCNA-low-1q-amp"). In the UkrAm dataset 7/84 cases (8%) showed loss on 22q which, however, never appeared "isolated" as described for a subgroup of tumors in the TCGA study [30]. Further, 19/84 (23%) cases showed gain of 1q in the UkrAm data set. The chromosomal bands 7q11.22-11.23, which contain the CLIP2 gene and was gained in genomic copy number in 29% of the UkrAm cases, was not present at all in the TCGA data subset. With regard to gene expression we performed unsupervised hierarchical clustering on the 31 UkrAm PTC cases (Figure SI 1) based on the mRNA expressions of a subset of 59 genes from the TCGA 71-gene signature. The missing 21 genes were not present in the Abend et al. (2012) expression data set [14]. The clustering revealed four main clusters C1, C2, C3, and C4 whilst notably all cases with a BRAF V600E mutation accumulated in cluster C3. No statistical association with any of the other tested parameters was found.

4.6 Integrative analysis on aCGH profiles and global mRNA expression data

Systematic integration of genomic copy number alterations with global gene expression did not reveal any significant associations between the copy number status of genes and corresponding expression at the mRNA level in the available data set.

5 Discussion

Our study is the first reporting on genomic CNAs in PTCs of the UkrAm cohort that integrates the findings with published data on global mRNA expression from a subset of the herein investigated UkrAm cases (Abend et al., 2012). Our exploratory omics approach aimed at the identification of molecular subtypes reflected by genomic copy number alterations that correspond to clinical features and most importantly to the radiation-associated parameters dose and latency. Several previous studies concluded that ionizing radiation potentially alters gene expression as a result of CNAs and subsequently induces the carcinogenic process, even at comparably low doses of radiation [34]. In a study by [9] that compared genomic copy number profiles of PTCs from patients who were exposed at very young age with PTCs from age-matched non-exposed patients the first radiation-specific CNA, gain of the chromosomal bands 7q11.22-11.23, has been reported . Further, a significant overexpression of CLIP2 which is encoded on 7q11.23 in exposed compared to non-exposed cases at the mRNA and protein levels has been demonstrated [9, 16]. Of note, the sets of tumors investigated in these studies differ with respect to mean age at exposure (approx. 2 years Hess et al. and approx. 8 years UkrAm), mean age at operation (approx. 16 years Hess et al. and 25 years UkrAm), and average dose (approx. 0.15 Gy Hess et al. and 0.68 Gy UkrAm). This lead to markedly different estimates of the proportions of radiation-induced PTCs in the exposed set of the Hess et al. (2011) study (approx. 85%) and the PTCs of the UkrAm cohort (55-75%) [3, 9, 16, 6]. Consequently, we hypothesized and also confirmed a lower frequency of gain of 7q11.22-7q11.23 in the UkrAm PTCs (29%) compared to the exposed set of PTCs of the Hess et al. (2011) study (39%). In a subset of the published TCGA cohort of sporadic PTC [30] that we matched to the herein analyzed cohort with regard to age the gain was not present

at all. This supports the role of gain of 7q11.22-11.23 as a radiation-specific marker since so far it was detected in radiation-associated PTCs only. Classification of the UkrAm PTCs into radiation-induced or sporadic cases based on an integrated analysis of CLIP2 protein expression and CLIP2 genomic copy number revealed a frequency of 75% radiation-induced cases which was in line with the above mentioned estimated frequency of 55-75% of radiation-induced cases [16]. The TCGA study on sporadic PTCs [30] already illustrated the dominant and mutually exclusive role of driver genetic alterations in PTC, in particular in the MAPK and PI3K pathways. Interestingly, the frequency of the BRAF V600E mutation in our investigated UkrAm cases was only 13% (9 out of 70 cases) and thus was significantly lower compared to the frequency of "BRAF-like" cases of the age-matched TCGA subset (62.8%). Likewise sporadic PTCs MAPK signaling is also deregulated in radiation-associated PTCs [17], most frequently by either rearrangement of the RET gene (referred to as RET/PTC) or by mutation of the BRAF gene (referred to as BRAF V600E mutation) in a mutually exclusive manner. In our study we found RET/PTC1 or RET/PTC3 rearrangements with a frequency of 31% (22 out of 71 cases) which is in the lower range of earlier reported frequencies of RET/PTC in young adults. In all, the observed frequencies of RET rearrangements and BRAF V600E mutations in our study were in good agreement with the findings on PTCs of the UkrAm cohort reported by Leeman-Neil et al. (2013) [17] (15% BRAF V600E mutated cases and 35% RET/PTC positive cases). In contrast the age-matched TCGA [30] PTC subset showed a much lower frequency of RET/PTC rearrangements (13.3%) compared to UkrAm which is likely to be due to differences in the distribution of histological subtypes in the different tumor sets since RET/PTC and BRAF V600E have been associated with particular histological subtypes or predominant histological structures [35]. Unfortunately, Agrawal et al. (2014) [30] only differentiated "classical-type", "follicular-variant", "tall cell variant" and "uncommon PTC variants" whereas our PTCs were classified using more detailed subtype categories, which did not allow a one-to-one comparison. Another interesting result of the large-scale analysis by the TCGA consortium is the molecular sub-classification of PTC into "BRAF-like" and "RAS-like" PTCs using the so called "BRAF-RAS score". Since the "BRAF-RAS score" is derived from a 71-gene-signature, which was determined by differential expression analysis of tumors harboring either BRAF V600E or RAS mutations, we hypothesized that this signature is capable to discriminate BRAF-mutated cases and RAS-mutated cases to a certain extent. The published signature has been derived from RNAseq data, whilst the UkrAm expression data originate from gene expression array data (Abend et al., 2012). This might explain that the Abend et al. array data lack some genes since in contrast to RNAseq microarrays cannot reliably detect genes that are weakly expressed. However, unsupervised hierarchical clustering based on the remaining 59 genes resulted in four distinct clusters and, consistently, all cases with a BRAF V600E mutation were present in one single cluster (C1) which reflects the "BRAF-like" cases. Hence, although we were not able to calculate a BRAF-RAS score due to the fact we did not have RAS mutation data available we could prove the usefulness of the underlying gene expression signature for the discrimination of PTCs into BRAF-like and non-BRAF like. Moreover, this finding suggests that the BRAF V600E driven molecu-

lar phenotype is well reflected by the "BRAF-like" associated genes. Concerning a comparison of the genomic copy number alterations of the PTCs investigated in our study and that of the age-matched subset of the Agrawal et al. (2014) [30] study we observed marked differences. 83% of the TCGA subset did not show any CNA at all, referred to as "SCNA-quiet" group in the Agrawal et al. (2014) paper [30]), whereas only 53% of the UkrAm PTCs (cluster C2.1 Figure 2 can be attributed to this group, although none of the UkrAm PTCs did not show no CNA at all, but only a few smaller ones. None of the other groups i.e. "SCNA-22q-del" and "SCNA-low-1q-amp" were found in our data set. This is likely to be due to the fact that the subset of the Agrawal et al. (2014) [30] we used for comparison was purely sporadic and, moreover, differed from our data set with regard to histological subtypes. With regard to the global genomic copy number unsupervised hierarchical clustering of the of the CNA profiles of the 84 UkrAm PTCs resulted in two main clusters with significantly more females, cases without lymph node metastasis and long latency (> 17 years) in cluster C1. This partly reflects the findings reported by Hess et al. [9] who described a similar result for CNA associations with the lymph node status [36, 29, 37]. Interestingly, the groups as they were defined by unsupervised hierarchical clustering can also be built considering only copy number changes of chromosomes 1q and 19. Hence, these alterations are the most discriminating ones amongst all CNAs. Chromosome 1q harbors, among many others, the genes TMP3 and TRIM33 (synonym RFG7) that have already been shown to be associated with PTC. While in radiation-associated PTCs TRIM33 has been shown to be fused to the RET gene as part of the RET/PTC7 rearrangement [38], TMP3 has been identified as a fusion partner of the NTRK1 gene in PTC [39]. In addition to these exemplarily mentioned thyroid cancer associated genes one can speculate that more genes located on altered chromosomes 1q and 19 may play an important role in driving radiation-induced PTC. Therefore, we suggest an in-depth integrated analysis of these genomic regions particularly with respect to mRNA and protein expressions of the affected genes. The systematic integration analysis of the genomic copy number level with the transcriptome level that we carried out in this study unfortunately did not result in any positive findings which might, on the one hand, be due to the limited number of cases that overlapped between our array CGH and the Abend et al. (2012) expression microarray data set [14]. On the other hand, it also might be an effect caused by intratumoral heterogeneity since the nucleic acids used in array CGH and expression microarray analysis were extracted from different parts of the tumors. Many of the CNAs from this study have already been described for PTC in young patients previously [9, 40, 41]. Stein et al. [40] reported in post-Chernobyl childhood PTCs chromosomal gains on 1q, 12q, 22q and a loss on 21q in post-Chernobyl childhood PTCs. Also, Hess et al. (2011) and Unger et al. (2008) presented overlapping CNAs [9, 41] with this study of the UkrAm cohort (gains on 1q, 9q, 16p/q, 19p/q, 20, 21q and loss on 11p). In UkrAm cases CNAs could be associated with latency > 17 years, absent lymph node metastasis, and female gender. Association testing also revealed a region on chromosome 22q (region no. 429, includes NF2 and CHEK2), significantly associated with latency. 35% of all cases with a latency > 17 years harbored the 22q gain, compared to 6% of cases with latency < 17 years. This result was even more prominent when only females were

considered for whom 46% of cases with latency $>$ 17 years and only 4% of cases with latency $<$ 17 years showed the gain. This represent an important observation in terms of radiation exposure of the UkrAm cases which is further complemented by the fact that [42, 29, 43, 14] copy number gains of two genes, FAM38A and MTA1 were associated with radiation dose. Of special interest these two genes were amongst the 11 genes that showed a statistical significant association of gene expression with radiation dose in the Abend et al. [14] study. Both genes showed a significantly higher expression in the intermediate dose group (300-1000 mGy) in the initial study. In the same dose group we found a significant enrichment of cases with the corresponding chromosomal gain confirming the findings by Abend et al. (2012) [14] at the genomic level and pointing to further radiation-associated genetic alterations in addition to gains on 7q11.22-11.23 and associated CLIP2 alterations. In conclusion, we present the first study on CNAs of PTCs from the UkrAm cohort and show that the carcinogenesis reflecting genomic landscape of these tumors is very heterogeneous. With regard to gain of 7q11.22-11.23 which was identified as a radiation-specific copy number alteration in PTC in a previous study on radiation-associated PTC (9), we observed a comparable frequency in our data set. This let us conclude that the subset of UkrAm cases showing the gain are similar to the radiation-induced PTCs from the previous Ukrainian cohort and also are likely to be radiation-induced [9]. Additionally, we considered findings that were published in a comprehensive multi-level omics study on a large set of sporadic PTCs suggesting that a subset of cases from UkrAm cohort is also of sporadic origin.

6 Acknowledgements

We thank the International Pathology Panel of the Chernobyl Tissue Bank for confirmation of diagnosis: Professors A Abrosimov, G Fadda, G Hant, V LiVolsi, J Rosai and ED Williams; the Chernobyl Tissue Bank for collection of thyroid tissue samples; Professor G Thomas as coordinator of the Chernobyl Tissue Bank; Professor Y Nikiforov for the exchange of BRAF V600E and RET/PTC data; Aaron Selmeier for technical support. This study was supported by the European Commission, EpiRadBio project, FP7 Grant No. 269553 and in part by the European Commission, DoReMi project, Grant No. 249689. The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/.

References

[1] Cardis, E. et al. (2011). The Chernobyl accident–an epidemiological perspective. *Clin Oncol (R Coll Radiol)*, **23**, 4, 251–60. ISSN 1433-2981 (Electronic) 0936-6555 (Linking). doi:10.1016/j.clon.2011.01.510.

[2] Ron, E. (2007). Thyroid cancer incidence among people living in areas contaminated by radiation from the Chernobyl accident. *Health Phys*, **93**, 5, 502–11. ISSN 0017-9078 (Print) 0017-9078 (Linking). doi:10.1097/01.HP.0000279018.93081.29.

[3] Brenner, A.V., et al. (2011). I-131 dose response for incident thyroid cancers in Ukraine related to the Chornobyl accident. *Environmental Health Perspectives*, **119**, 7, 933–939. ISSN 1552-9924 (Electronic) 0091-6765 (Linking). doi:10.1289/ehp.1002674.

[4] Cardis, E., et al. (2005). Risk of thyroid cancer after exposure to 131I in childhood. *J Natl Cancer Inst*, **97**, 10, 724–32. ISSN 0027-8874. doi:10.1093/jnci/dji129.

[5] Davis, S., et al. (2004). Risk of Thyroid Cancer in the Bryansk Oblast of the Russian Federation after the Chernobyl Power Station Accident. *Radiation Research*, **162**, 241–248.

[6] Tronko, M.D., et al. (2006). A cohort study of thyroid cancer and other thyroid diseases after the chornobyl accident: thyroid cancer in Ukraine detected during first screening. *J Natl Cancer Inst*, **98**, 13, 897–903. ISSN 0027-8874. doi:10.1093/jnci/djj244.

[7] Zablotska, L.B., et al. (2011). Thyroid cancer risk in Belarus among children and adolescents exposed to radioiodine after the Chornobyl accident. *Br J Cancer*, **104**, 1, 181–7. ISSN 0007-0920. doi:10.1038/sj.bjc.6605967.

[8] Unger, K., et al. (2004). Heterogeneity in the distribution of RET/PTC rearrangements within individual post-Chernobyl papillary thyroid carcinomas. *The Journal of clinical endocrinology and metabolism*, **89**, 9, 4272–4279. ISSN 0021-972X. doi:10.1210/jc.2003-031870.

[9] Hess, J., et al. (2011). Gain of chromosome band 7q11 in papillary thyroid carcinomas of young patients is associated with exposure to low-dose irradiation. *Proc Natl Acad Sci U S A*, **108**, 23, 9595–600. ISSN 0027-8424. doi:10.1073/pnas.1017137108.

[10] Bozhok, Y., et al. (2006). A Cohort Study of Thyroid Cancer and Other Thyroid Diseases After the Chornobyl Accident. Cyto-Histopathologic Correlation and Accuracy of Fine Needle Aspiration Biopsy in Nodules Detected During the First Screening in Ukraine (1998-2000). *Cancer*, **107**, 11, 2559–2566. ISSN 0008-543X. doi:10.1002/cncr.22321.

[11] Tronko, M., et al. (2012). Thyroid cancer in Ukraine after the Chernobyl accident (in the framework of the Ukraine-US Thyroid Project). *J Radiol Prot*, **32**, 1, N65–9. ISSN 1361-6498 (Electronic) 0952-4746 (Linking). doi:10.1088/0952-4746/32/1/N65.

[12] Likhtarev, I., et al. (2006). Questionnaire- and measurement-based individual thyroid doses in Ukraine resulting from the Chornobyl nuclear reactor accident. *Radiat Res*, **166**, 1 Pt 2, 271–86. ISSN 0033-7587 (Print) 0033-7587. doi:10.1667/rr3545.1.

[13] Likhtarov, I., et al. (2005). Post-Chornobyl Thyroid Cancers in Ukraine. Report 1: Estimation of Thyroid Doses. *Radiation Research*, **163**, 2, 125–136. ISSN 0033-7587. doi:10.1667/RR3291.

[14] Abend, M., et al. (2012). Iodine-131 dose dependent gene expression in thyroid cancers and corresponding normal tissues following the Chernobyl accident. *PLoS One*, **7**, 7, e39103. ISSN 1932-6203 (Electronic) 1932-6203 (Linking). doi:10.1371/journal.pone.0039103. Abend, Michael Pfeiffer, Ruth M Ruf, Christian Hatch, Maureen Bogdanova, Tetiana I Tronko, Mykola D Riecke, Armin Hartmann, Julia Meineke, Viktor Boukheris, Houda Sigurdson, Alice J Mabuchi, Kiyohiko Brenner, Alina V eng Randomized Controlled Trial Research Support, N.I.H., Intramural Research Support, Non-U.S. Gov't 2012/08/01 06:00 PLoS One. 2012;7(7):e39103. doi: 10.1371/journal.pone.0039103. Epub 2012 Jul 25.

[15] Abend, M., et al. (2013). Iodine-131 dose-dependent gene expression: alterations in both normal and tumour thyroid tissues of post-Chernobyl thyroid cancers. *Br J Cancer*, **109**, 8, 2286–94. ISSN 0007-0920. doi:10.1038/bjc.2013.574.

[16] Selmansberger, M., et al. (2014). CLIP2 as radiation biomarker in papillary thyroid carcinoma. *Oncogene*. ISSN 1476-5594. doi:10.1038/onc.2014.311.

[17] Leeman-Neill, R.J., et al. (2013). RET/PTC and PAX8/PPARgamma chromosomal rearrangements in post-Chernobyl thyroid cancer and their association with iodine-131 radiation dose and other characteristics. *Cancer*, **119**, 10, 1792–1799. ISSN 1097-0142. doi: 10.1002/cncr.27893.

[18] Buffart, T.E., et al. (2007). DNA quality assessment for array CGH by isothermal whole genome amplification. *Cell Oncol*, **29**, 4, 351–9. ISSN 1570-5870 (Print) 1570-5870 (Linking).

[19] Gentleman, R.C., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*, **5**, 10, R80. ISSN 1465-6914 (Electronic) 1465-6906 (Linking). doi:10.1186/gb-2004-5-10-r80.

[20] Neuvial, P., et al. (2006). Spatial normalization of array-CGH data. *BMC Bioinformatics*, **7**, 264. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi:10.1186/1471-2105-7-264.

[21] Sjoerd Vosse, M.v.d.W. (2009). CGHregions: Dimension Reduction for Array CGH Data with Minimal Information Loss. *R package version 152*.

[22] Wiel, S.V.a.M.v.d. (2012). CGHbase: Base functions and classes for arrayCGH data analysis. *R package version 1200*.

[23] Vucur, M., et al. (2013). RIP3 inhibits inflammatory hepatocarcinogenesis but promotes cholestasis by controlling caspase-8- and JNK-dependent compensatory cell proliferation. *Cell Reports*, **4**, 4, 776–790. ISSN 2211-1247. doi:10.1016/j.celrep.2013.07.035.

[24] van de Wiel, M.A., et al. (2005). CGHMultiArray: exact P-values for multi-array comparative genomic hybridization data. *Bioinformatics*, **21**, 14, 3193–4. ISSN 1367-4803 (Print) 1367-4803 (Linking). doi:10.1093/bioinformatics/bti489.

[25] Selmansberger, M., et al. (2015). Dose-dependent expression of CLIP2 in post-Chernobyl papillary thyroid carcinomas. *Carcinogenesis*. ISSN 1460-2180. doi:10.1093/carcin/bgv043.

[26] McCullagh, P. et al. (1989). Generalized Linear Models, Second Edition. doi:SBN-13: 978-0412317606,ISBN-10:0412317605.

[27] Benjamini, Y. et al. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B (Methodological)*, **57**, 1, 289–300. ISSN 00359246. doi:10.2307/2346101.

[28] Likhtarov, I., et al. (2006). Post-Chernobyl thyroid cancers in Ukraine. Report 2: risk analysis. *Radiat Res*, **166**, 2, 375–86. ISSN 0033-7587 (Print) 0033-7587. doi:10.1667/rr3593.1.

[29] Cancer Genome Atlas Research, N. (2014). Integrated genomic characterization of papillary thyroid carcinoma. *Cell*, **159**, 3, 676–90. ISSN 1097-4172 (Electronic) 0092-8674 (Linking). doi:10.1016/j.cell.2014.09.050.

[30] Agrawal, N., et al. (2014). Integrated Genomic Characterization of Papillary Thyroid Carcinoma. *Cell*, **159**, 3, 676–690. ISSN 0092-8674. doi:10.1016/j.cell.2014.09.050.

[31] Gao, J., et al. (2013). Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science Signaling*, **6**, 269, pl1. ISSN 1937-9145. doi:10.1126/scisignal.2004088.

[32] van Wieringen, W.N., et al. (2012). Matching of array CGH and gene expression microarray features for the purpose of integrative genomic analyses. *BMC Bioinformatics*, **13**, 80. ISSN 1471-2105 (Electronic) 1471-2105 (Linking). doi:10.1186/1471-2105-13-80.

[33] van Wieringen, W.N. et al. (2009). Nonparametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, **65**, 1, 19–29. ISSN 1541-0420 (Electronic) 0006-341X (Linking). doi:10.1111/j.1541-0420.2008.01052.x.

[34] Mullenders, L., et al. (2009). Assessing cancer risks of low-dose radiation. *Nat Rev Cancer*, **9**, 8, 596–604. ISSN 1474-1768 (Electronic) 1474-175X (Linking). doi:10.1038/nrc2677.

[35] Santoro, M., et al. (2006). RET/PTC activation in papillary thyroid carcinoma: European Journal of Endocrinology Prize Lecture. *European Journal of Endocrinology*, **155**, 5, 645–653. ISSN 0804-4643, 1479-683X. doi:10.1530/eje.1.02289.

[36] Alfthan, K., et al. (2004). Cyclic AMP-dependent protein kinase phosphorylates merlin at serine 518 independently of p21-activated kinase and promotes merlin-ezrin heterodimerization. *J Biol Chem*, **279**, 18, 18559–66. ISSN 0021-9258 (Print) 0021-9258 (Linking). doi:10.1074/ jbc.M313916200.

[37] Hartmann, M., et al. (2015). Tumor Suppressor Nf2 Blocks Cellular Migration by Inhibit- ing Ectodomain Cleavage of Cd44. *Mol Cancer Res*. ISSN 1557-3125 (Electronic) 1541-7786 (Linking). doi:10.1158/1541-7786.MCR-15-0020-T.

[38] Rabes, H.M. (2001). Gene rearrangements in radiation-induced thyroid carcinogenesis. *Med Pediatr Oncol*, **36**, 5, 574–82. ISSN 0098-1532 (Print) 0098-1532 (Linking). doi:10.1002/mpo. 1133.

[39] Greco, A., et al. (2010). Rearrangements of NTRK1 gene in papillary thyroid carcinoma. *Mol Cell Endocrinol*, **321**, 1, 44–9. ISSN 1872-8057 (Electronic) 0303-7207 (Linking). doi: 10.1016/j.mce.2009.10.009.

[40] Stein, L., et al. (2010). Copy number and gene expression alterations in radiation-induced papillary thyroid carcinoma from chernobyl pediatric patients. *Thyroid*, **20**, 5, 475–87. ISSN 1050-7256. doi:10.1089/thy.2009.0008.

[41] Unger, K., et al. (2008). Array CGH demonstrates characteristic aberration signatures in human papillary thyroid carcinomas governed by RET/PTC. *Oncogene*, **27**, 33, 4592–602. ISSN 1476-5594 (Electronic) 0950-9232 (Linking). doi:10.1038/onc.2008.99.

[42] Wreesmann, V.B., et al. (2004). Genome-wide profiling of papillary thyroid cancer identifies MUC1 as an independent prognostic marker. *Cancer Res*, **64**, 11, 3780–9. ISSN 0008-5472 (Print) 0008-5472 (Linking). doi:10.1158/0008-5472.CAN-03-1460.

[43] Lambros, M.B., et al. (2010). PPM1D gene amplification and overexpression in breast cancer: a qRT-PCR and chromogenic in situ hybridization study. *Mod Pathol*, **23**, 10, 1334–45. ISSN 1530-0285 (Electronic) 0893-3952 (Linking). doi:10.1038/modpathol.2010.121.

Table 1: main features of patients

	n cases		AaE	AaO	Latency		Dose
female	51	**mean [yrs]**	8.4	25.7	17.4	**arithmetic mean [Gy]**	0.95
male	33	**median [yrs]**	9.0	26.0	16.6	**geomeric mean [Gy]**	0.32
Total	84						

Table 2: CNAs associated with the paramaters latency, gender and dose

Test parameter	Groups	n	Gains	Losses
Latency	$<$ 17 years	47	3	3
	$\geq$ 17 years	37	174	36
Gender	female	51	73	1
	male	33	-	-
Latency (Females)	$<$ 17 years	27	3	-
	$\geq$ 17 years	24	162	45
Dose	$\leq$ 300 mGy	40	-	-
	301-1000 mGy	18	28	-
	$>$ 1000 mGy	26	-	-

Figure 1: Frequency plot of CNAs of the PTCs investigated in our study. Green bars starting from the top indicate the percentage of CNA profiles with copy number gains in the data set and red bars from the bottom indicate the percentage of CNA profiles with copy number losses in the data set.

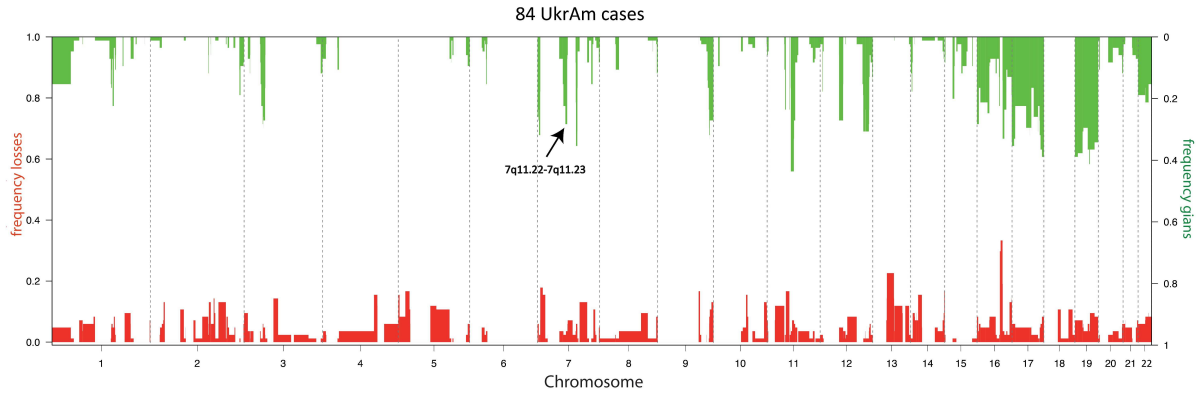

Figure 2: Unsupervised hierarchical cluster analysis of CNA profiles resulting in a heatmap corresponding to the clustered CNA profiles (left, copy number gains are depicted in blue and losses in red). The status of the parameters gender, lymph node status, latency, AaO, AaE, CLIP2, BRAF V600E, RET/PTC1/3 and dose and indicated by colored bars right of the heatmap. The clustering dendrogram is shown on the right.